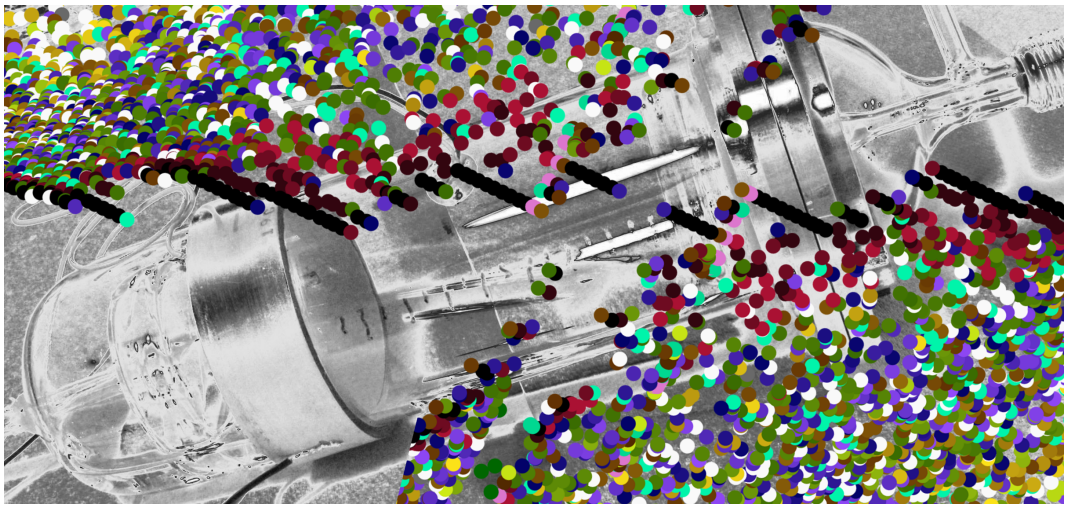
Design and optimisation of acoustic resonators for sonofusion experiments

Thesis accepted in partial fulfillment of the requirements for the degree
DOKTOR DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

by the KIT-Department of Mechanical Engineering of the
Karlsruhe Institute of Technology (KIT)

by

Dipl.-Phys. Markus Julius Stokmaier



completed at the Institute for Nuclear and Energy Technologies (IKET),
Karlsruhe Institute of Technology (KIT)

Day of oral exam: May 6th 2019
Advisor: Prof. Dr.-Ing. habil. Andreas G. Class
Co-advisor: Prof. Dr.-Ing. Martin Gabi

Karlsruhe, April 2020

Design and optimisation of acoustic resonators for sonofusion experiments

(Design und Optimierung akustischer Resonatoren für Sonofusionsexperimente)

Zur Erlangung des akademischen Grades eines
DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

von der KIT-Fakultät für Maschinenbau des
Karlsruher Instituts für Technologie (KIT)
angenommene

DISSERTATION

von

Dipl.-Phys. Markus Julius Stokmaier

Tag der mündlichen Prüfung: 6. Mai 2019
Hauptreferent: Prof. Dr.-Ing. habil. Andreas G. Class
Korreferent: Prof. Dr.-Ing. Martin Gabi

Acknowledgements

This thesis was elaborated while working as research associate at the Institute for Nuclear and Energy Technologies (Institut für Kern- und Energietechnik, IKET) of the Karlsruhe Institute of Technology (Karlsruher Institut für Technologie, KIT) which, in 2019, was renamed Institute for Thermal Energy Technology and Safety (Institut für Thermische Energietechnik und Sicherheit, ITES). The preceding work of resonator characterisation documented in the appendix chapters was carried out while being delegated by the KIT as visiting scholar to the Gaerttner Laboratory of the Rensselaer Polytechnic Institute (RPI).

Firstly, I would like to express my deep gratitude to Prof. Richard T. Lahey, the initiator of the RPI-KIT sonofusion project who is a tireless supporter, to my advisor Prof. Dr.-Ing. habil. Andreas Class who contributes so many ideas, thought associations, and enthusiasm to support and push forward in every situation, and to Prof. Dr.-Ing. Thomas Schulenberg who is excellent in always being a steady supporter and sharp critic at the same time. Above all, I want to thank you three for being great teachers and discussion partners. For their interest, discussion-partnering, and critical thinking I am deeply grateful to Prof. Dr.-Ing. Xu Cheng and to the late Prof. Dr.-Ing. Martin Gabi who had taken over the co-advisorship.

Dr. Lahey, I also want to thank you for having welcomed me in Troy, NY, and within your group of students, and for having introduced me into the RPI Linac team. In the group I met Silvina Cancelos, who set the ground for simulation-based resonator optimisation at RPI and helped so much in quickly lessening my low-orientation newcomer status in the group, on campus, and in the lab. The lab stand for resonator characterisation was the core of the research project preceding the simulation-focused work. Without Bernie Malouin's initial setup work and without him as co-worker during the multiple dis- and re-assembly runs leading to understanding of the setup and the data, my time would have been a much harder time, and a more monotonous one. Similarly, I don't want to imagine a shared Linac office without Frank Saglime and Rian Bahran as mates, always there with critical thought when you need it, and ready to discuss anything related to nuclear physics and the rest of the universe. Frank was the builder and conductor behind Dr. Lahey's RPI sonofusion experiment campaign and he investigated countless resonators. He's a great explorer of neutron physics & programmable math wizardry, and also great company for exploration around the globe. All you, I cannot thank enough.

Uphill of Troy's Rensselaer campus, you wouldn't assume you're next to an almost historic accelerator lab when you drive by the Gaerttner Lab's office building half grown in behind thick green bushes. Since the late 50s the Lab hosts a massive pulsed neutron source with really neat time-of-flight measurement specs and no need to be hidden at all. Within, you find the best of exceptional American research drive and innovation spirit, but you know that only after having had the chance to hear Peter Brand's red plastic "That was easy!" buzzer button a couple of times, have your electronics ideas straightened by his counter-questions, and have your mechanics construction trials straightened out with help from Matt Gray, Martin Strock, and Azzeddine Kerdoun. When they don't teach thread drilling to visiting scholars they busy themselves with playing an accelerator control room of a size able to fit the lunar rover and maybe even a bit more classy in its steampunk looks, where slight infusions of silicon valley modernity are not allowed to endanger the overall style. Former head Robert Block and current director Yaron Danon are overseeing and artfully organising this extraordinary site of continuous collaboration, creativity, and fundametal research. I feel very grateful that you folks up at the Linac have allowed me to be one of the team.

A special thanks goes to Devin Barry, another member of the Linac team. Without his insightful argumentation for commencing with data analysis and plotting scripts in Python I would never have gotten so quickly so deeply into programming. It is doubtful whether I would have started experimenting with evolutionary algorithms of my own design without him.

The IKET has been shaped by the continuous work of its former director Thomas Schulenberg and group leaders like Andreas Class, who managed to establish a research culture of openness, curiousity, mutual support, and creating advantage from perfectly interlocking methods and competences. It seems not always to be the norm. Among my teachers, coworkers, and friends, I can count many who have practised and lived this culture and spirit in this organisation and who have enriched it every day by a highly disputative central agora, the Kaffeepause. For helpfulness every day and years of insightful and witty debating I want to thank all of you. Special thanks goes to Tino Ortega Gómez for continuous motivating company and critical interest in the sonofusion resonator topic. I owe very special thanks to Manuel Raqué, dauntless fighter for style and refinement, not only in scientific writing. I feel honoured, having enjoyed your help and company as office mate for several, however, much too short years.

For his artful picture and visualisation contributions I thank Tudor Pirvu. For their email advice for realising the displacement measurement with a turntable pickup needle I thank Michael Pettersen from Shure and Carsten Lindegaard Anderson from Ortofon. For info, data, and advice for modelling piezoelectric material I thank Steve Poterala and Brian Nava from Channel Industries. For their helpful answers to my questions about APDL coding and best practice hints on FE modelling and model evaluation I want to thank Slav Dimitrov and Martin Hanke of Cadfem GmbH. For much appreciated discussions about evolutionary algorithms and the properties of the CMA-ES in particular I want to thank Niko Hansen from INRIA (Institut national de recherche en sciences et technologies du numérique). For boosting my understanding of plasma physics and the MAGO experiment with detailed email answers I want to thank Sergey Garanin from VNIIEF. For sharing beautiful pictures of plasma in a Farnsworth Hirsch fusor I want to thank Thomas Rapp of Rapp Instruments.

For all their continuous support during long years of research I am deeply grateful to my parents and my sister. Mom, Dad, I thank you both for having brought us into so much contact with nature and for having always encouraged and answered our questions about origins and reasons.

Weimar, April 2020 Markus Stokmaier

Abstract

In 2002 publications by Taleyarkhan et al. claiming successful sonofusion (SF) [456–459] started a scientific dispute which has remained unsettled up to the present day. It motivated the staging of several replication trials, but their ambiguous results could not clarify the situation. This means, the question whether sonoluminescence (SL) in imploding cavitation bubbles yields a type of plasma generation mechanism which can also lead to thermonuclear fusion plasma is still awaiting a proper answer. The fact that the physics and thermodynamics of the imploding bubble as an extremely transient multi-scale phenomenon is not completely understood theoretically certainly contributes to the unresolved state of the topic.

In SF trials according to Taleyarkhan et al. an acoustic resonator of the design by West & Howlett [504, 505] is used which exhibits both in experiments and FEM simulations an extreme degree of sensitivity with respect to working point and design parameters. As long as the manufacturing of the main components involves manual glassblowing, the sensitivity can hardly be minimised. The present work is based on the hypothesis that the mechanics of this type of acoustic resonator were not examined and understood well enough in the past, and that this is a central reason for the persisting ambiguous state of affairs around SF even after several replication trials. In order to fill this gap, detailed parametrised finite element simulation models of the West-Howlett type resonator were developed in preceding projects. In the present work, new acoustic resonator geometries are proposed and investigated numerically for their ability to produce similar sound field patterns and their suitability for future SF experiments. A crucial precondition for the investigated design proposals is a substantial reduction in tolerances based on the transition to part machining techniques enabling better precision.

Simple resonators like organ pipes, laser devices, or microwave generators are similar in the aspect that the boundary conditions are rigidly fixed for an oscillating field quantity within a bounded inner volume. By contrast, liquid-filled resonators for cavitation experiments feature a strong coupling between the inner fluid and the structural hull, they might as well be called acoustic-mechanic resonators. The resonances of such a resonator are determined by the interplay of acoustic and mechanic degrees of freedom. In order to tune such a system to yield a maximal internal acoustic pressure and to enable the highly energetic collapse of acoustic bubbles, it is crucial to balance out the different field spaces, masses, and springs. The result is a multi-dimensional optimisation problem exhibiting a multitude of local optima arising from a competition of manifold resonances. It is an important consequence that a comparison between differing resonator geometry concepts can only be meaningful

under the condition that of each design variant a thoroughly optimised instance is made available.

For this reason, evolutionary algorithms (EA) as efficient global search techniques are investigated for their potential to solve the problem of the parametric optimisation of resonator designs. Based on literature, the EA branches of evolution strategies with covariance matrix adaptation (CMA-ES), particle swarm optimisation (PSO), and the concept of EA hybridisation are identified as promising. The most suitable workflow for the task is systematically derived via performance benchmarks on a deliberately selected collection of test functions representing key features of the envisaged application. A specially developed hybrid EA turns out to be very efficient with respect to the set of benchmark problems.

Ultimately, this work provides an objective comparison of three newly developed resonator geometries with the hitherto employed West-Howlett design. Two design proposals prove to be well-performing whereas the simplest geometry reveals as its deficiency the lack of necessary vibrational degrees of freedom. In total, a substantially improved knowledge and methodology base is being established because:

- the design space of non-trivial resonator geometries can only be efficiently explored with the help of simulations,
- the objective comparison of design variants can only be achieved with an effective optimisation workflow of global search and local tuning,
- only the transition to more precise manufacturing techniques allows (a) the straightforward implementation of designs developed by simulation and (b) to reach a state of reproducibility of the discussed type of SF experiments.
- A considerable improvement lies also in the fact that the establishment of a fast, standardised, effective optimisation methodology allows a faster feedback between experiment and simulation. This enables tailored re-adjustments of resonator designs or the simulation model itself (“calibration”) upon obtaining new experimental data and generally renders the computer-aided backing of experimental campaigns much more agile.

The aim is to improve the pre-conditions for future SF experiments decisively. The proposal of new resonator designs which can be precision-machined to match well-characterised working points can achieve this because it opens up a way towards reproducible experiment conditions. The well-founded prospect of reproducibility with optimised equipment can serve as justification for undertaking renewed efforts to demand from nature an answer to the question of sonofusion, i. e. the question whether the phenomenon of SL can be extrapolated to achieve fusion-capable plasma.

Zusammenfassung

Seit den Veröffentlichungen von Taleyarkhan et al. zur Sonofusion (SF) [456–459] und den dadurch motivierten Replikationsversuchen ist der Forschungsstand ungeklärt verblieben. Das heißt, die Frage, ob Sonolumineszenz (SL) in implodierenden Kavitationsblasen eine Art der Plasmaerzeugung ist, die auch einen Weg zu fusionsfähigen Plasmazuständen liefert, harrt nach wie vor einer eindeutigen Antwort. Dies liegt nicht zuletzt daran, dass auch von theoretischer Seite her die Physik und Thermodynamik der kollabierenden Blase als extrem transientes Multiskalenphänomen nicht bis ins Detail verstanden ist.

Bei den SF-Versuchen nach Taleyarkhan et al. wird ein akustischer Resonator vom Design nach West & Howlett [504, 505] verwendet, der sich im Experiment und der FEM-Simulation als extrem sensitiv auf Betriebs- und Designparameter zeigt. Die hohe Sensitivität lässt sich bei Verwendung der Herstellungstechnik der Glasbläserei auch nur schwer einschränken. Die vorliegende Arbeit basiert auf der Hypothese, dass die nicht genügend untersuchte und verstandene Funktionsweise dieses Typs eines akustischen Resonators ein zentraler Grund für den widersprüchlichen Gesamtstatus der SF-Replikationsversuche ist. Davon ausgehend wurden in vorangegangenen Arbeiten Finite-Elemente-Rechenmodelle des West-Howlett-Resonatortyps erstellt, um diese Lücke zu füllen. In der hier vorliegenden Arbeit werden nun neue Designvarianten analog konzipierter akustischer Resonatoren vorgeschlagen und auf ihre Eignung für zukünftige SF-Experimente hin untersucht. Grundvoraussetzung der untersuchten Designvorschläge ist hierbei eine deutliche Verringerung der Maßtoleranzen durch den Übergang zu präziseren Herstellungstechniken.

Einfache Resonatoren wie Orgelpfeifen, Lasergeneratoren und Mikrowellengeräte ähneln sich insofern, dass die Randbedingungen fest sind und eine innere Feldgröße oszilliert. Im Gegensatz hierzu zeichnen sich flüssigkeitsgefüllte Resonatoren für Experimente mit akustischer Kavitation dadurch aus, dass das innere Fluid und die äußere Struktur einer starken Kopplung unterliegen, weshalb man auch von akustisch-mechanischen Resonatoren sprechen kann. Die Resonanzen eines solchen Resonators werden durch das Zusammenspiel von akustischen und mechanischen Freiheitsgraden bestimmt. Soll das System auf maximalen Schalldruck im Inneren des Resonators optimiert werden, um einen möglichst energiereichen Kollaps von Kavitationsblasen zu ermöglichen, müssen die Schallfeldräume, Massen und Federkonstanten des Resonatordesigns aufeinander abgestimmt werden: es entsteht ein vieldimensionales Optimierungsproblem, das aufgrund eines Wettbewerbs verschiedener Resonanzen viele lokale Optima aufweist. Eine zentrale Konsequenz ist, dass ein Vergleich mehrerer Resonatordesignansätze nur dann aussagekräftig ist, wenn

von den Designvarianten gründlich optimierte Instanzen vorliegen.

Aus diesem Grund werden evolutionäre Algorithmen (EA) als effiziente globale Sucher zur Lösung des Problems der parametrischen Resonatordesignoptimierung evaluiert. Aus der Literatur werden die EA-Varianten der Evolutionsstrategie mit Covarianzmatrixadaptation (CMA-ES) und der Partikelschwarmoptimierung (PSO) sowie der Ansatz der EA-Hybridisierung als vielversprechend identifiziert. Der beste Ablauf zur Resonatoroptimierung ergibt sich durch eine Performanzmessung anhand bewusst ausgewählter Testfunktionen, die Schlüsseleigenschaften der Anwendung repräsentieren. Ein eigens entwickelter hybrider EA erweist sich als sehr leistungsfähig in Bezug auf die Testfunktionsauswahl.

Schlussendlich liefert diese Arbeit einen objektiven Vergleich dreier neuentwickelter alternativer Resonatorgeometrien mit dem bisher verwendeten West-Howlett-Design. Im Ergebnis erweisen sich zwei der drei untersuchten Designvorschläge als performant, während sich bei der einfachsten Geometrie als klares Manko ein Mangel an nötigen Schwingungsfreiheitsgraden identifizieren lässt. Es wird somit eine deutlich verbesserte Wissens- und Methodikgrundlage für zukünftige SF-Experimente geschaffen, denn

- nur simulationsgestützt lässt sich der Designraum nichttrivialer Resonatorgeometrien effizient durchforsten,
- nur ein effektiver Optimierungsablauf aus globaler Suche und lokalem Feintuning erlaubt den objektiven Vergleich von Designvarianten,
- nur der Umstieg zu präziseren Herstellungstechniken erlaubt (a) die gezielte Umsetzung rechnergestützt entwickelter Designs und (b) ganz grundsätzlich das Erreichen reproduzierbarer SF-Versuchsbedingungen.
- Schließlich erlaubt das Etablieren einer schnell und standardisiert einsetzbaren Optimierungsmethodik auch das bedarfsgerechte Nachjustieren von Designs sowie des eigentlichen Simulationsmodells (“Kalibrierung”) nach Erlangung neuer experimenteller Befunde; die rechnergestütze Begleitung experimenteller Kampagnen wird entscheidend agiler.

Somit will diese Arbeit die Grundbedingungen für zukünftige Sonofusionsexperimente entscheidend verbessern. Neue präzise gefertigte und wohlcharakterisierte Designs, wie sie in dieser Arbeit vorgeschlagen werden, versprechen reproduzierbare Versuchsbedingungen. Die begründete Aussicht auf reproduzierbare Versuche unter optimierten Bedingungen liefert eine Rechtfertigung, die Frage, ob Sonolumineszenz einen Weg zu thermonuklearer Fusion erlaubt, erneut an die Natur zu stellen.

Contents

Chapter 1

Introduction & context

This chapter gives a short overview on the history of sonofusion (SF) trials spawned by the research on sonoluminescence (SL). It outlines the principles underlying SL as a phenomenon of extraordinary energy concentration and how this has led to diverse attempts of developing SF devices. After a spotlight on the controversy which has unfolded around the SF trials of Taleyarkhan et al., the last sections concentrate on the SF research campaign at RPI and the contribution of the collaborative RPI-KIT research efforts which started with experimental resonator characterisations and later focused on developing improved finite element models of the acoustic resonators.

1.1 A claim and a controversy

In 2002 a group of researchers around Taleyarkhan et al. published an extraordinary claim in an edition of Science magazine [458]: to have achieved thermonuclear fusion in a purely compression-heated plasma in a table-top lab setup involving not much more than a glass vessel, a piezoelectric transducer, a liquid containing deuterium (see figure 1.1), and some standard electronic equipment. The table-top fusion claim was based on collapsing acoustic cavitation bubbles, which are known to have an extraordinary capability of energy concentration and can yield hot, light-emitting plasma. The group made headlines with "bubble fusion" and upheld the claim in later publications [325, 456, 457, 459, 463, 466]. Soon however, there were smaller and larger problems shading the seeming success story, the biggest one was (and still is) that this small-scale fusion experiment did not turn out to be promptly reproducible by other research teams in other labs.

The observing fellow scientists and journalists explored many different theories for explaining the ambiguous status, scientific and non-scientific narratives. This work starts with the question whether an alternative explanation of the situation may be found in the high sensitivity and unreliable performance of a central piece of equipment, the acoustic resonator, and it intends to progress the stalling dispute by suggesting better resonator equipment with more predictable and reliable properties.

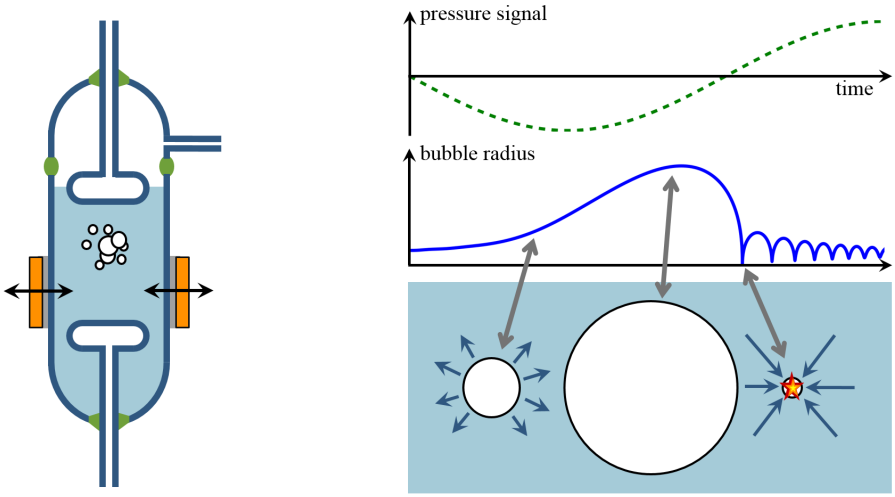


Figure 1.1 A resonator for acoustic cavitation and sonofusion
The sketch on the left shows the basic cross section geometry of the acoustic resonator used by Taleyarkhan et al. for their bubble fusion experiments. The piezoelectric transducer (orange) drives the vibration motion and indirectly the sound pressure field inside the liquid-filled resonator. The goal is to create bursts of cavitation bubbles which grow at low pressure and collapse violently at high pressure. The exemplary bubble size history on the right shows a particularly important feature of bubble dynamics: the slope on the contraction side is much steeper, i. e. the interface speed of the collapsing bubble can become very fast and the unavoidable turnaround very harsh. The result is the transfer of a large amount of kinetic energy from the fast-moving liquid onto the bubble's internal gas content which serves as the spring. That this compression can lead to incandescent plasma is known since many decades, the phenomenon is called sonoluminescence. If it becomes possible to achieve hot enough plasma to enable nuclear fusion, then it can be called sonofusion.

1.2 Root and motivation of the present resonator optimisation study

Directly after the first publications of Taleyarkhan et al. there was a great interest to replicate the extraordinary table-top fusion experiment based on the simple ingredients of a glass resonator and a little bit of electronics. R. T. Lahey Jr., who was among the authors of the Science piece and later articles, initiated an experimental campaign of replication trials at Rensselaer Polytechnic Institute (RPI) in Troy, NY. The team started with resonator parts received from Taleyarkhan's group and continued developing own resonator variants. The work resulted in a master thesis and three reports [68, 250, 252, 390] without yielding the hoped-for bubble fusion confirmation. What these reports do not show is the series of different resonator exemplars which were manufactured, characterised, and – in the case of the exemplars with the most promising properties – tested for their capabilities of generating cavitation bubbles with fusion plasma, all with negative results except one seemingly positive signature [390][1] which stayed however below the threshold of statistical significance.

This was when the focus of the RPI team shifted to the resonator design and when the hardly reproducible resonator properties coupled with instable and shifting working points were identified as a main obstacle preventing a steady conduct of productive experimental work. While most of the literature dispute on bubble

[1] in particular figures 67 and 69 on pages 92 and 93

fusion revolved around the means of tracing fusion (radiation detection and nuclear physics), they realised that the engineering problem of achieving a more robust resonator design might lie prominently in the way in front of the nuclear instrumentation questions. In cooperation with Cancelos a new type of resonator design was developed from scratch and validated in 2D-axis-symmetric finite element simulations with slightly simplified geometries [69]. The results of the prototype tests were however disappointing: an all-glass version cracked due to an inherent problem of stress distribution and a glass-aluminium composite version brought the liquid inside not to good cavitation conditions but instead to the boiling point because of an excessive amount of vibration energy dissipation.

Around this time Lahey had established a cooperation with the Institute of Nuclear and Energy Technologies (IKET) of the Karlsruhe Institute of Technology (KIT), and after a first experimental campaign, this cooperation focused on better finite element (FE) simulation models of the acoustic resonators and ultimately on the task of exploring new design ideas to find solutions to the encountered problems.

One of the first investigations conducted with the improved FE model was a sensitivity analysis yielding two strong insights. The main result was a proof of the high sensitivity of the resonator performance with respect to many parameters determining design, position tuning of parts, and the working point. It confirmed many disappointing practical experiences made before. The first insight was that without changes in the techniques for manufacturing and setting up the resonator systems, no condition of reproducibility would be easily achievable. The consequent next step was to further pursue the work begun by Cancelos and to seek new resonator designs allowing the desired change in techniques and enabling a better control of manufacturing precision and working conditions.

When exploring new designs, a second insight from the sensitivity study comes into play: an objective comparison between two resonators cannot be made if one or both are more or less mistuned, an objective judgement can only be made between two well-tuned, i. e. thoroughly optimised, design instances. At this point the motivation was given to find an efficient way to algorithmically tune resonator models. This thesis describes the simulation-aided exploration of newly developed resonator design ideas, and above all the development and application of a suitable algorithmic design optimisation workflow involving global search with the help of evolutionary algorithms.

1.3 From sonoluminescence to sonofusion

Chemistry is part of our everyday experience and can be as easy to handle as lighting a match. But nuclear reactions need sophisticated machinery like accelerators and huge reactor buildings to be controlled. Modern physics made the understanding of both things possible and was able to reveal the fundamental similarity. In chemical as in nuclear reactions particles are being re-grouped in different ways: atoms are re-grouped into different molecules in chemical reactions whereas in nuclear reactions protons and neutrons are re-ordered. Physics also explains why the two types of reactions are so different, why the particle energies in nuclear reactions are many orders of magnitude higher, why machinery for controlling nuclear reactions like

accelerators and nuclear reactors came so much later in history as compared to mankind mastering fire.

Sonoluminescence (SL) is light emission from a tiny volume of an almost adiabatically heated plasma inside a quickly imploding gas bubble in a liquid. It is a common side effect of acoustic cavitation and has been well documented since the early 20$^{\text{th}}$ century. The standard laboratory equipment for creating sonoluminescence is very simple and consists of a piezoelectrically actuated glass resonator filled with liquid. Sonoluminescence is the consequence of a very efficient energy concentration mechanism in which the spherical symmetry of an imploding bubble, stabilised by surface tension, allows to reach the high energy densities of chemical reactions going out from the low energy density of the sound field in the liquid filling the acoustic resonator. Could it be possible to modify the setup of an SL experiment in such a way as to further increase the energy density achieved in the tiny plasma volume sufficiently so that thermonuclear fusion reactions could occur? In that case *sonofusion (SF)* would become observable in addition to sonoluminescence. Sonofusion is a very interesting topic because it might drastically lower the technological threshold necessary for achieving controlled nuclear fusion reactions. However, the reproducible experimental proof of sonofusion (or *"bubble fusion"*) would definitely not automatically open up a straightforward way to harvest energy.

The claimed experimental observations of sonofusion by Taleyarkhan et al. in 2002 at ORNL[2] [458] started a major debate in the scientific community. However, to date no completely independent group of scientists has been able to reproduce the ORNL experiment successfully. This work examines one possibility for explaining the current research status on the topic of sonofusion: could the sensitivity of the resonator design that was used be the crucial factor hampering the experiment's reproducibility? The answer is deemed to be yes, triggering the next question: could it be overcome? Therefore it is attempted to add to the debate about bubble fusion by investigating the resonator design problem.

Before directly addressing the analysis of the resonator design used by Taleyarkhan et al. and the issue of how to develop a more performant and better reproducible design in the main chapters, a short introductory part will outline the physics and some of the historical context of the sonofusion question. Moreover, additional extended context descriptions are made available in appendix chapters A & E. Since the topic of sonofusion has created a wider interest in the past, some of these texts are aimed at a broader readership.

1.3.1 What is sonoluminescence?

Sonoluminescence means light generated with sound [199, 284]. It can be observed as a side effect of acoustic cavitation in liquids. It is light emitted from tiny volumes of highly compressed and almost adiabatically heated plasma [85, 96, 205, 246]. When cavitation bubbles implode, the kinetic energy of the moving liquid is transferred onto the compressed gas within the shrinking bubble where it raises the pressure and temperature. The light emission pulses imply that the temperature becomes high

[2]Oak Ridge National Laboratory, Oak Ridge, TN, USA

enough so that collisions among the atoms lead to the stripping of valence electrons and the dissociation of chemical bonds [187, 449, 529].

Cavitation

When gas is in contact with a solid, then every gas particle being reflected by the surface of the solid transmits a certain amount of momentum. The time- and space-averaged effect of many gas particle collisions is a force per unit surface area: pressure [14]. The pressure can be lowered by making the gas particles slower and sparser, it can become zero in vacuum, but it cannot become lower than that.

In a liquid, however, where the attractive forces between particles dominate their motion and determine the macroscopic properties of the ensemble, there can of course exist negative forces across test surfaces, one speaks of states of tension [76, 260]. A seeming contradiction can arise when considering that above any free liquid surface a finite positive vapour pressure can be measured [14]. How can a liquid bear a state of tension and prevent the rupture of its continuum at a pressure below the level which would immediately be measurable inside a newly created cavity? The question can be resolved by taking into account surface tension. It requires an offset pressure inside any bubble in order to stabilise it. The surface tension-induced offset pressure[3] grows with the curvature of the bubble surface and becomes very large for tiny bubbles. Therefore, a liquid can be ruptured either if the state of tension is large enough to directly overcome the inter-particle potential wells or, at lower tension, if bubbles larger than a critical limit are supplied as seed, so that the difference between the vapour pressure and the external tension is larger than the surface tension-induced pressure offset [56].

This means that cavitation always has to be discussed considering the degree of purity of the liquid. Behind a ship propeller cavitation occurs as soon as centrifugal forces of spinning water lower the pressure at the centre of vortices below the vapour pressure. This is because micron-sized seed bubbles are omnipresent in seawater and because any small pressure drop immediately creates new seed bubbles if a high level of dissolved gasses is present in the liquid. Strong states of tension are thus only possible in highly purified liquids [56, 260]. In the context of SL and SF this is relevant because its creation in a liquid under strong tension allows a bubble to accumulate more potential energy during its lifecycle. This happens through the increase of the expansion speed, whereby more inertia of the surrounding liquid moving outwards leads to a larger final radius during the rarification cycle, before the bubble collapse during compression. How exactly these interacting forces of external & internal pressure, surface tension as well as liquid inertia play out in determining the radial motion of the bubble interface is described by the *Rayleigh-Plesset equation*.

The Rayleigh-Plesset equation

Due to its compressibility a gas bubble in a liquid can be excited into oscillatory radial motion. The gas may be considered to be the spring and the surrounding

[3]The pressure inside a bubble of radius R is $\Delta p = 2\sigma/R$ higher compared to its environment, where σ is the surface tension.

liquid the mass. The oscillation can be excited by an external harmonic sound field. The equation of motion ($F = ma$) of this system is the Rayleigh-Plesset (RP) equation [56, 257, 259, 276]. A short explanation and outline of a derivation of the RP equation can be found in appendix C. The RP equation produces highly nonlinear effects which can be seen in the simulated time series depicted in figures 1.2 and 1.3.

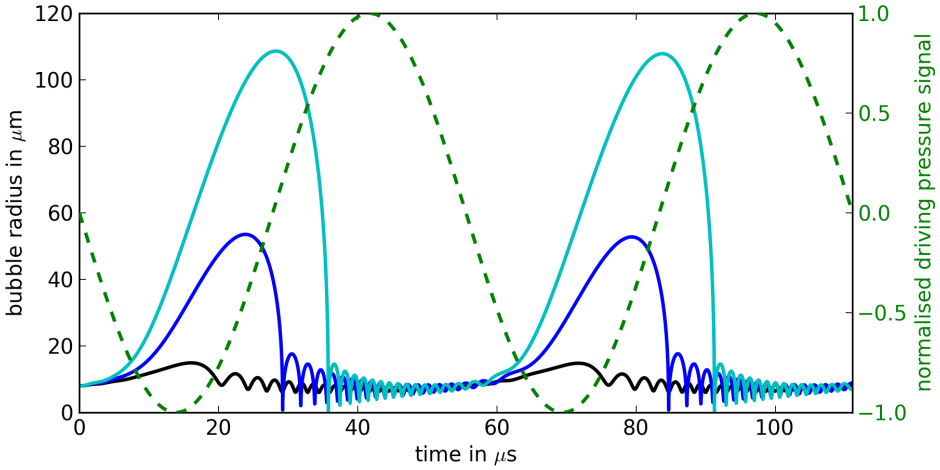


Figure 1.2 Simulation of an oscillating bubble
The diagram shows solutions of the Rayleigh-Plesset equation for a bubble in acetone with an equilibrium radius R_0 of 8 µm driven by a sinusoidal external sound field at 18 kHz. The computations were made solely for illustrative purposes (the code can be found at [431]). The three traces in black, blue, and cyan correspond to different and increasing acoustic pressure levels of 0.85, 1.1, and 1.4 bar. Three important consequences of the increasing sound pressure level can be seen: it leads to (a) a larger bubble size as initial condition for the collapse, to (b) a delayed time of collapse under higher external pressure due to inertia, and consequentially to (c) higher final speeds (5.6, 2.1×10^3, and 22×10^3 m s^{-1}) of the bubble interface.

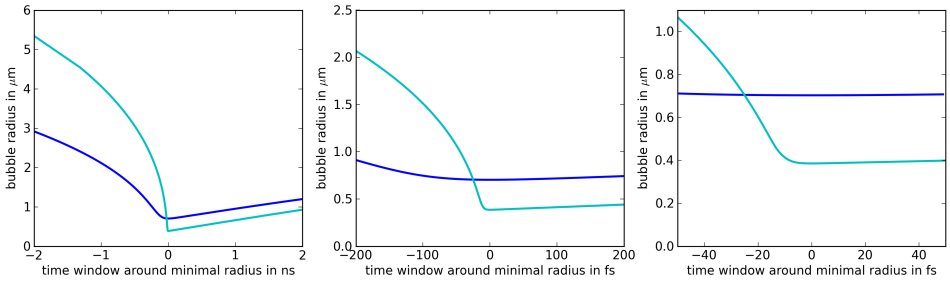


Figure 1.3 Close-up around the moment of highest acceleration
The above plots show details of the same time series as displayed in figure 1.2. On the one hand they reveal how the coarser time resolution hides drastic speed increases in the late phases of the collapse. On the other hand the differences between incident and emergent angles illustrate the energy loss due to sound radiation.

The time-histories show the periodic response of a bubble to an external sinusoidal pressure excitation. A steady bubble growth is initiated when the external pressure is low. The faster the maximal speed of growth, the longer is the duration of the growth period. This is the simple consequence of inertia. A noteworthy feature is the steep slope on the contraction side after the peak. The lines tend to become almost vertical and this is why the differences in collapse speed of two of the three

curves can only be visualised through extremely zoomed-in close-ups as depicted in figure 1.3. The continuation of the fast collapse until a harsh turnaround is again the consequence of inertia. The most important consequence in our context is the energy concentration mechanism, the transfer of large amounts of kinetic energy from the fast moving mass, the surrounding liquid, onto the spring, the gaseous bubble content made up by a relatively small amount of particles. If the compression is strong and fast enough, then energy dissipation by heat transfer mechanisms will be too slow and the temperature rise inside the bubble can become sufficient to turn the gas into an incandescent plasma.

Speed, curvature, and the Mach number

The Mach number $M = v/c$ is the ratio between the local fluid velocity and the speed of sound and can often be taken as an indicator for the transition between different physical regimes. In the case of imploding cavitation bubbles the low-Mach regime ($\dot{R} \ll c$) means that there are no sharp pressure gradients inside or outside the bubble. But when the interface speed $\dot{R}$ approaches, or exceeds, the speed of sound in the gas or liquid phase the situation changes and gradients or even shock waves will arise. Concerning shock waves in the liquid, direct experimental evidence is available in the form of photographies of shock fronts emanating from rebounding bubbles [257, 345]. As for the gas inside the bubble, the situation is not so clear. In the absence of direct experimental observations, the indirect way using numerical simulations and their calibration is the only route so far allowing a description of pressure and density profiles and their transient evolution. In the high-Mach regime the pressure and density right in front of the converging interface will rise faster than farther ahead towards the bubble's centre. One important question is whether the density wave piling up in front of the interface can turn into a shock front and whether this shock front can detach, accelerate, and leave the interface behind. The speed of a shock wave increases with the amplitude of the pressure jump. Together with the increasing curvature of a concentric spherical wave front, this can explain the principle idea of a shock front forming and beginning to run ahead of the interface during the terminal phase of bubble implosion. However, whether simulations allow shock formation or not, depends on many model assumptions and boundary conditions, which is the reason why few conclusive answers are available for the many different types of SL experiments [85, 257]. In any scenario involving a converging shock front (of cylindrical, spherical, or intermediate ellipsoidal shape), it represents additional stages of energy concentration going beyond the bubble itself, due to the rapid increase in speed and amplitude of the shock as it approaches and impacts on itself at the centre of the imploding gas bubble.

Single- and multi-bubble sonoluminescence

Cavitation and sonoluminescence can be induced in liquids in many different ways. Cavitation can be understood as 'rupturing the liquid' when fast transient elastic motion patterns raise the tension in the liquid to the level where the inter-particle forces forming the liquid are overcome. There are no cracks in a liquid. Wherever cavities appear, their existence lowers the tension instantaneously in their vicinity

(the pressure inside being ≥ 0), so that inter-particle forces dominate the particle ensemble again and surface tension as bubble-shaping force is re-established. This turns what would otherwise be cracks into veils and filaments of myriads of microbubbles populating the central regions of expanding motion patterns. If it's not just a pulsed signal but oscillatory cycles of an acoustic pressure field, then forces of bubble interaction arise reorganising bubble populations. Exemplary photographies of resulting filament structures can be found e. g. in [4, 298]. When SL arises during the collapse of such clustered bubbles, then it is called *multi-bubble sonoluminescence (MBSL)*. As there are many ways of creating cavitation bubble clusters by pulsed shocks or oscillating fields, based on seeded or nucleated bubbles, or simple "rupture", there are many regimes of MBSL to be examined.

In contrast to MBSL the term *single-bubble sonoluminescence (SBSL)* has been coined for describing a very different and particular experimental setup where one single bubble is captured within the antinode of an acoustic sound field and set into radially oscillating motion.[4] The time-averaged force resulting from the interaction of the oscillatory bubble motion with the external sound field is called the *Bjerknes force*. The fact that the Bjerknes force allows levitation of a sonoluminescing bubble at the centre of a resonator is very advantageous if one wants to focus with optical instruments on SL as a light source. Another important aspect of SBSL is that the time-development of the bubble content is brought to a stationary state over the course of many oscillations as gasses coming out of solution replace[5] the vapour and inert gasses replace molecular species. The latter happens because the products of dissociation reactions in the SL plasma tend to enter the solution state in the liquid. Thus, SBSL in water, without purification and if no other gasses are artificially added, stems effectively from an argon bubble in water [85].

1.3.2 SL: spectroscopy, interpretation, models

SL spectra and their interpretation

SL generally has two contributions: a continuum and overlaying emission band structures. Not every report on SL observations describes both parts, but there are SBSL as well as MBSL setups in which both contributions can be observed [449]. Emission band structures can be interpreted more directly and easily. Firstly, the existence of excited states hints towards energies reached and information like the degree of ionisation. Secondly, if the assumption of a Boltzmann distribution of populated energy levels is valid, then the comparison of relative peak intensities allows a direct deduc-

[4] The experimental technique of SBSL has been discovered and rediscovered several times, by Yosioka & Omura in 1962 [531], by Temple in 1970 [468], and finally Crum & Gaitan [150, 152] in the late 1980s who managed to popularise it.

[5] Two important drivers influencing the bubble's gas content are the area and the shell effect [85, 260]. The area effect creates a net inflow of gas because times of low bubble pressure mostly overlap with times of large interface area. The shell effect has to do with gas concentration gradients in the liquid and can be understood by imagining what happens to surrounding shells of liquid during bubble expansion and shrinking. While expansion leads to a thinning of layers of liquid which increases any radial concentration gradient the opposite happens during contraction. As a result, diffusion is accelerated during bubble expansion which coincides mostly with times of low pressure and slowed during contraction; this leads again to a net inflow of gas.

tion of the temperature (*two-line radiance ratio method* [448]). Furthermore, looking at spectral lines with sufficiently fine frequency resolution allows the interpretation of shape details of the peaks.

In a noteworthy article [133] Flannigan & Suslick investigate the frequency-shifted, broadened, and asymmetrically deformed peaks of Argon emission lines from SBSL in concentrated sulphuric acid. In the context of a quantum-mechanical treatment of the Stark effect the peak-broadening (*pressure [or Stark] broadening*) can be understood as the consequence of radiating particles seeing strong electric fields from neighbouring particles and thus becomes a diagnostic tool to measure the plasma density with little temperature sensitivity [125]. The second order Stark effect introduces asymmetries in the peak broadening and allows a clearer interpretation than the broadening itself because of no interference with other effects. For sound pressure amplitudes between 2.7 and 3.8 bar they deduced temperatures ranging from 7000 to 16 000 K and electron densities from 4×10^{17} to 4×10^{21} electrons per cubic centimetre. This maximal plasma electron density N_e, the authors remark referring to [386], was "comparable to that generated by the Lawrence Livermore National Laboratory Nova laser (1.8 kJ in 1 ns at 527 nm) in inertial confinement fusion experiments on a polyethylene target."

The example of the peak-broadening by the Stark effect shows exemplarily how data interpretation can hinge on the assumed theoretical model of the physics at play. This is what makes the interpretation of the other contribution to SL spectra, the underlying continuum, extremely difficult and ambiguous. Molecular emissions like in a candle flame can create a continuum but bremsstrahlung may as well [449]. Several continuum-producing photon emission processes employed in SL models are described in some detail in [187]. They comprise blackbody radiation, electron-ion bremsstrahlung, recombination radiation, electron-neutral atom bremsstrahlung, and collision-induced emission (i. e. collisions of neutral atoms at lower temperatures). Generally, combinations of these processes need to be considered. What complicates the task of finding and fitting SL models, is that not only internal plasma properties like species, temperatures, or mean free paths have to be approximated, but also external boundary conditions like when and how the initially isothermal bubble contraction transitions into an adiabatic compression, rates of condensation and evaporation at the interface, and so on. Articles like [188, 204, 312, 489] have been cited in reviews [85, 449] as exemplary successful model fits.

SL models

A model of SL is a multi-stage theory beginning with some form of the Rayleigh-Plesset equation, dealing with the continuum mechanics and the energy household (chemical and thermodynamic) inside the bubble, and ending with the qualitative and quantitative analysis of various radiation emission processes. The main questions are:

- Is the compression-heated volume inside the bubble in thermal equilibrium or does the bubble centre become hotter than the rest (hot spot model, compression wave model)?

- Can a shock front develop instead of just a compression wave?
- What chemical species are present in the gas phase and at which ratios? (There is rectification over repeated collapses but the ratios are also changing during one single cycle due to evaporation and condensation.)
- What amount of energy is consumed by the dissociation of molecules?
- What are the plasma conditions (T, p, ionisation degree, ...)?
- Which light/radiation emission mechanisms are relevant?
- Are there specific reactions which can be incorporated into the model and fitted to emission band structures?
- How transparent is the plasma or does it have an opaque core (i. e. ratio of photon mean free path and plasma size, volume or surface emission)?
- Transient nature: what is the time- and space-dependence of the conditions in the gas phase?
- Transient nature: is it necessary to treat electron and ion gasses separately?
- MBSL: which different scenarios are encountered by different bubbles in a cluster [291, 326, 448]?

After more than two decades of intensified research following the popularisation of SBSL the field has grown quite large. An overview can be gained via books like [85, 94, 532], review articles [26, 57, 84, 187, 257, 368, 449, 501, 527, 529, 530] (most recent: [95] of 2015), and a look at some noteworthy recent articles [113, 133, 448].

Thinking about the possibility of SF, the question of shock waves deserves special consideration. Generally, the formation of a concentric shock wave in the gas phase would have two important consequences. It would mean an additional energy focusing mechanism on top of the bubble motion itself with the potential to raise the peak temperatures by orders of magnitudes. On the other hand the concentration of energy in a small mass fraction of the bubbles gas content means a much smaller size of the plasma core and much less fusion reactions which might occur in comparison with a scenario where most of the bubble content except perhaps an outer layer comprises the hottest plasma.

The idea that perhaps the bubble content cannot be treated as homogeneous any more during the final stages of collapse, and that this thought can be extrapolated to the postulation of *microshocks*, dates back to the work of Jarman of 1960 [221]. Along with him, the review authors of [26, 57] list Löfstedt et al. [275], Greenspan & Nadim [180], Wu & Roberts [520], Barber et al. [28], and Moss et al. [310] as the inventors and developers of the first shock wave models. Shock-based models are a way to explain SL pulse widths in the range of a few picoseconds. It has been pointed out in review articles outlining the development of SL research [57, 85, 187] that during a time period in the mid-90s reports on extremely short pulse widths (<50 ps [27] or even <12 ps [308]) seemed to challenge SL models. A shift was brought about by the seminal publication by Gompf et al. [174] of the University

of Stuttgart who introduced the measurement technique of *time-correlated single photon counting (TCSPC)* to SL research. By basing the time measurement on the relative time delay of two identical photomultipliers, they could take much of the physics of the photon detection machinery out of the equation.[6] According to the reviews [57, 85, 187] it had since then been widely accepted that SL pulse widths range from 50 to >250 ps. Brenner [57] points out that a "ballpark" estimate can be made by comparing the time of light emission with the time scale of a sound wave traversing the bubble. He makes an example calculation with a bubble size of one micron and a speed of sound of $1000\,\mathrm{m\,s^{-1}}$ yielding an acoustic time scale of one nanosecond. It means that the shortest SL pulses of 50 picoseconds are 1.5 orders of magnitude away from the regime where this quick estimate would indicate a largely homogeneous bubble interior. According to this estimate it does not seem improbable that there are some SL bubbles for which an isotropic model could mean missing a decisive part of the description. It also suggests to pay attention to the speed of sound in the gas phase and makes experiments interesting in which that parameter can be controlled (e. g. by the molar weight of the gas).

Historic and alternative theories of sonoluminescence

A compression-heated incandescent gas is not the only explanation that has been put forward to explain sonoluminescence. Since the discovery of SL in 1934[7] by Frenzel & Schultes [147] many theories came up, and some are still being discussed. Frenzel & Schultes themselves assumed discharge after charge separation by friction due to the bubbles' motion through the liquid [370]. Before precise timing measurements were available, it was not sure whether the light emission occurred during bubble birth or collapse. This is why there were also interpretations of SL as a form of triboluminescence, the luminescence observable when some crystals are crushed. According to Cheeke [84], also in the line of electric discharge theories both options were discussed: (a) amplification of potential gradients across the liquid-gas interface during compression (*balloelectric theory* of SL) or (b) amplification of random charge fluctuations during the expansion phase. Some more details and citations about these early theories can be found in the review article by Cheeke [84].

In fact, in these early days of SL research, hot spot models were the more exotic ones [85, 95]. In 1950 Noltingk & Neppiras put forward the idea that compression-heated gas could be the origin of SL [327]. A few years later, Jarman [221] was the first to suggest the existence of concentric compression waves or microshocks [85, 95].

The perhaps most exotic theory explains SL as a quantum electrodynamic (QED) effect due to virtual photon pairs of the QED vacuum becoming real photons under

[6] Much of the physics does not mean everything. With TCSPC one can get to the dozens of picoseconds range with photomultipliers based on the detection principle of electron avalanches which take from 10^1 to $>10^2$ ns to build up (transit time). But there is a machine- and setup-specific transit time spread. TCSPC cannot resolve pulse shapes narrower than the transit time spread [218, 350, 494].

[7] The blackening of photographic plates as consequence of SL had been observed even before by Marinesco & Trillat in 1933 [286] but they misinterpreted their results as ultrasound having acted directly on the chemicals on the plate.

perturbation when an interface separating vacuum from a dielectric moves quickly [121, 302, 408]. This would mean SL glows similarly as black holes which are assumed to emit *Hawking radiation* in the form of a blackbody spectrum when single photons from virtual pairs are swallowed at the event horizon. The theory was discussed for a while in the 1990s ([302] contains a list of citations outlining the dispute) but more and more detailed calculations showed in the end that the theory couldn't hold up because either the bubble interface had to move faster than the speed of light or the radiation power would be by many orders of magnitude too weak.

According to the compilation by Crum [95], the list of more recent alternative theories includes emitting electrons confined in voids in compressed and hot liquid, confined electrons in nonhomogeneous compressed helium gas, proton-tunneling when water undergoes phase changes, or fracturing of liquid when jets after having traversed the bubble strike liquid again (*fractoluminescence/triboluminescence*).

Also still appearing on that list are publications on electric discharge theories of SL [261, 284, 285]. Since discharge models have been associated with SL since its discovery, it is the most persistent group of alternative theories. The fundamental idea is charge separation during asymmetric bubble or droplet deformation or breakup. Right behind the surface of water or other liquids there can be an electrical double layer of lopsided charge distributions. There is a difference in mobility because the first layer is adsorbed while the second, the shielding layer is diffusive. The fast-changing geometries of needle-like jets protruding into the bubble or of bottlenecks when daughter bubbles break off a larger bubble can be imagined to lead to substantial charge separation and potential gradients. Discharge theories of SL assume that these gradients become large enough to enable breakdown and the creation of discharge plasma in the gas or the liquid.

Another recent alternative explanation comes from Dey & Aubry [115] who attribute the light emission to the compressed liquid inside the supersonic shock front going out from rebounding bubbles after having almost reached the Van-der-Waals hard core radius.

With respect to these alternative theories of SL Brenner et al. [57] note: "In fact, all nonthermal models have to explain why their mechanism of emission would not be swamped by thermal radiation."

SL research: unconventional ways to approach the topic

Rounding off this review of the topic of SL, three interesting and particularly original, inventive, and illustrative recent research works will be outlined below: laser-manipulation of SL plasma, molecular dynamics simulations showing mass segregation, and plasma size calculations via interference with dust particles.

Khalid, Kappus et al. [232] of Putterman's UCLA group managed to produce extremely bright SL in a particularly large single bubble inside a quartz vial slammed repeatedly against a wall. With a laser ray they test the opacity of the SL plasma, not by examining how much laser light reaches the other side (a signal which would also be influenced by refraction and more difficult to interpret) but simply by using a strong enough laser with the effect that if the plasma is opaque enough to absorb a substantial fraction of the laser light, that then the change in plasma conditions due

to the additional energy input doesn't go unnoticed. With this technique, by poking it with a laser, Khalid et al. circumvented the usual problem that SL plasma cannot be examined with a probe. One of the main results is imagery showing that the plasma is so dense that the additional heating occurs in a very asymmetric manner concentrating on the side of the incident laser ray.

In [30] Bass et al., who are also part of or associated with the same UCLA group, report on having observed mass segregation in the simulation of a collapsing bubble. They used molecular dynamics (MD) simulations for their numerical experiments which means going a decisive step below the macroscopically emergent world of continuum mechanics. The additional degrees of freedom enable to examine what a classical hydrodynamic simulation cannot show.

Pingpong balls are reflected by soccer balls and many or very fast pingpong balls are needed to change the direction of a soccer ball. This is why in thermal equilibrium particles of different weight have different velocity distributions. In the non-equilibrium scenario of a compression or shock wave the same physics leads to mass segregation. The converging heavier particles squeeze out the many times reflected smaller particles. In the case of a shock front, the gas of the small particles evaporates out in front of the shock at average velocities much higher than the motion of the heavier species forming the primary shock front. Bass et al. could thus observe besides the effect of mass segregation that the gas of light particles gets squeezed by the heavy-particle shock front coming from behind and acting as piston. Beginning with the same initial total energy this leads to much higher peak temperatures in the simulated plasma than comparison scenarios with just one particle species in the gas phase.

The third example is the work of J. S. Dam and M. T. Levinsen [99, 100] from the University of Copenhagen. The original plan was to measure the size of SL plasma through a Hanbury-Brown Twiss (HBT) experiment. The HBT technique is used to measure the size of stars and exploits the boson properties of photons tending to share quantum-mechanical wave functions (unlike fermions, e. g. electrons, which exclude each other from populating the same wave function (Pauli principle)). One can imagine that photons originating from different locations on the surface of a far away star, before being measured in one single detector at the same time, must travel for a long time in close proximity. Their wave function overlap leads to a measurable degree of correlation, detectable as coincidence in time, or bunching. For stars of a given distance the degree of correlation is stronger the smaller the star is and hence the average distance between trajectories. SL plasma offers shorter photon trajectories but is at the same time the much smaller light source which is how the idea came up to examine it through a HBT experiment. Although its realisation was not possible within the project as measurement times were estimated to be in the range of months in order to get meaningful HBT correlation signals, it led Dam & Levinson to discover a different size measurement technique triggered by what first appeared as an imperfection: namely dust particles floating by between the SL source and the detectors. Due to the wave nature of light, no matter how small a light source is or how sharp the edge of a shadow-throwing object, there is always an interference pattern extending slightly into the supposedly shadowed area. Such interference patterns even amplified the light intensity in the centre

of the shadow (as compared to outside the shadow) behind dust particles of the right size in their experiments. Dam & Levinson turned the careful interpretation of these interference patterns into a new size estimation technique for SL plasma. They concluded that the light-emitting region in their setup was significantly smaller than the bubble's minimal radius which would support the notion of non-homogeneous plasma conditions (i. e., spherical shock waves within the imploding bubbles).

1.3.3 SL and the question of sonofusion

Why is SL such a sensitive phenomenon so that on the theoretical as well as the experimental side the peak temperature is sometimes inferred as being a couple of thousand and sometimes as 100 000 Kelvin? One important reason for that sensitivity (and an important hint that there is indeed a question of sensitivity and not only uncertainty) lies in the spherical symmetry. The spherical symmetry of the collapsing bubble leads to a singularity in a mathematical description in terms of continuum mechanics, and the same would be true on the level of a converging shock in the gas phase if there is one. This results in sensitivity because small variations in setup and boundary conditions can have a large effect on how much the path towards the singularity is followed and on how much the peak is reduced by the imperfections of real-world scenarios.

This is why the potential to reach thermonuclear fusion conditions was associated with SL plasmas not long after the first hot spot theories of SL were elaborated (see e. g. [11, 26, 28, 50, 60, 96, 187, 308, 311]). Indeed, a patent [135] and publication [136] by Flynn bear witness that the idea of a cavitation bubble-based fusion reactor dates back at least to the late 70s.

Fusion energy as motivation

Nuclear fusion promises CO_2-free energy with practically endless resources and a much less severe radioactive waste problem compared to fission reactors. Flynn's cavitation fusion reactor (CFR) design is a simple symmetric box full of liquid lithium or beryllium with ultrasonic horns centred on each face so that sound waves focus on the centre of the cuboid liquid metal volume where seeded deuterium bubbles can be imploded. The assumption is that the "acoustically induced cavitation fusion" works while the liquid metal bath is kept hot. Under this condition fusion products can deposit their kinetic energy in the hot bath which allows the extraction of the fusion energy input by conventional means like heat exchangers and a Rankine steam cycle with turbines. At the same time this consideration should make it clear that in Taleyarkhan et al.'s SL setup where the liquid has to be kept cool preventing the build-up of too much vapour pressure in the bubble offers no viable way of efficient energy extraction even if there was a way to scale up to more than their miniscule fusion rates. So, in order to enable energy harvesting via SF:

- a liquid has to be found in which SF can be demonstrated at high temperatures allowing energy extraction from a hot bath,
- each bubble implosion has to trigger more than just a few fusion reactions, and

- it must be possible to scale-up the initial geometry to an array of active SF sites (and to enable the sustain of a chain reaction in case neutron-scattering is used as the means of bubble nucleation).

Therefore, work on a reproducible SF setup based on a single active site in a cooled organic liquid or aquæous solution is clearly fundamental research and any association with the idea of SF as a potential energy source is speculative.

Another aspect often mentioned in the motivational context of SF is *tabletop fusion*. It has to be noted that tabletop fusion is nothing new and nothing special any more. Causing deuterium and tritium to fuse by accelerating the particles with electric fields can be accomplished on a very small scale. The examples are commercially available and portable accelerator tubes serving as neutron generators (e. g. for imagery & detection purposes) or *pyro fusion* [320] where the accelerating field is created by pyroelectric crystals. The novelty of SF as a potentially new type of tabletop fusion source would just be associated to the fact that it would be tabletop thermonuclear fusion instead of tabletop accelerator-driven fusion. In the speculative context of SF as energy source this is meaningful because of the difference in entropy of the directional accelerator setup which doesn't allow net energy extraction and the high-entropy thermal plasma which would (in principle) allow it.

Further motivation for SF research

Imploding bubbles represent a remarkable way of energy concentration bridging many orders of magnitude from the acoustic sound field to plasma and opening up the research fields of sonochemistry and sonoluminescence. Sonochemistry is an interesting field reported on in dedicated scientific journals and it has important applications like the degradation of toxic chemicals. SL represents a remarkably simple way of experimenting with hot dense plasmas which are otherwise much harder to produce (e. g. laser confinement or magnetised target compression). As researchers around the globe have such a simple and cheap access to extreme plasma conditions at their disposal, the enhanced plurality boosts our understanding and progress of theoretical plasma models and also of various multi-scale physical phenomena belonging to the explanation of SL. As can be seen when comparing literature in the context of other (fusion) plasma confinement methods, each new plasma production and confinement method adds a new channel of knowledge gain and new distinct problems triggering the advancement of experimental techniques and theoretical models. Therefore, SL research has an important impact on plasma research as it offers new opportunities to check and extend our understanding of plasma physics with a particular focus on transient phenomena. This is true for SL with or without SF. However, if SF were to become reproducibly demonstratable, then the counting of fusion neutrons would also become a new and very directly interpretable data channel for quantitative analyses. When SL plasmas turn into SF plasmas, it would be the indication of a clear threshold being reached, a strong nail for fixing theoretical models to the reality. Considering the limits of sonochemistry (only the end products can be analysed after everything has cooled down) or spectrometry (from the bubble core to the detector absorption and refraction occurs in the plasma, at the

bubble's interface, in the surrounding liquid and the resonator or window materials), neutrons offer such a high-quality data channel because their scattering and absorption on the way to the detector can be modelled very precisely by well-understood physics. Therefore, in terms of fundamental research two basic motivations can be formulated: SF as an additional plasma research technique and SF as pushing the mastering of small-scale dense plasmas to a new level.

1.3.4 Different SF experiment setups reported in literature

The earliest documented experimental setup with the goal of achieving thermonuclear fusion inside a cavitation bubble seems to be Hugh Flynn's cavitation fusion reactor (CFR) [135, 136] mentioned above, a cuboid container for liquid metal with face-centred acoustic horns for focusing acoustic waves in the middle where cavitation bubbles are to be imploded. An interesting aspect about the patent is that two dedicated setups are described as fusion reactions are deemed to be possible both inside the bubble (CFR type II) and in a shell of liquid surrounding it (type I). In the type II setup a nozzle at the bottom of the volume furnishes deuterium and tritium-filled bubbles rising through the liquid volume. It is anticipated that due to gravity and the rising motion of the bubble its collapse may not be spherical and magnetic fields for cancelling gravitation forces are imagined as a counter-measure (however, without mentioning the implication of Lorentz forces on charge carriers in the liquid metal and their impact on the collapse motion). The working mode of type I aims at the fusion of dissolved deuterium and tritium occurring in the shell of liquid surrounding the bubble when that shell is compressed and heated due to the extremely abrupt speed reversal when the bubble rebounds. The elegance of this idea lies in the fact that the chemical composition of the liquid phase needs to be controlled instead of the composition of the gas species inside the bubble. The concentration of dissolved deuterium in the metal can be controlled through setting a certain gas pressure above a free surface of the liquid, whereas the bubble contents can be hard to control due to the transient nature of the cavitation bubble.

In 1990, an article in a Russian-language journal by Lipson et al. [272, 273] reported the observation of fusion neutrons accompanying cavitation in heavy water containing metal powder particles. The paper features a table of neutron count rates containing significantly elevated values for more than one experimental setup. Unfortunately, it seems that there have been no repetitions of this experiment and the scientific community failed to add much to the briefly enumerated and speculative ideas of explanation (including friction, fracture, and electric fields) in the original publication.

In the course of the 90s, after the popularisation of SBSL, the possibility of SF (or *bubble fusion*) suddenly became a more broadly discussed topic [28, 60, 311, 370]. It has to do with shock-based theories of SL [26, 520] having been put forward to explain extremely short measured SL pulse widths of around or below 50 ps [85, 187, 529]. As shock waves represent another energy focusing mechanism saddled on top of the bubble dynamics, these theories generally imply increased peak temperatures and pressures which can bring a theoretical model into the range of fusion conditions. Subsequently measured longer pulse widths gave more room for competing theories

not relying on shock waves [85, 187, 284]. Reported SBSL pulse widths of a few hundreds of picoseconds excluded shockwave-based theories as explanations for these experiments [57, 529]. Nevertheless, particular SL setups with observable short pulse widths down to about 50 ps remained part of the discussion [174, 206].[8]

At the end of that decade having witnessed the flourishing of many SL models, funding from DARPA[9] was supplied to several research groups with the particular goal of investigating the possibility of achieving sonofusion. Among the several research groups the controversial papers by Taleyarkhan et al. claimed successful SF observations [456, 458, 459]. were one result of the campaign. The setup described therein is based on bubble clusters in deuterated acetone nucleated by external neutrons and will be discussed in more detail below. An investigation of the corresponding acoustic resonator technology used by Taleyarkhan et al. is the primary focus of this dissertation.

In the years since then there have been several attempts to replicate that particular SF experiment, some published in journals [67, 414], one[10] even through TV. All fully independent groups reported negative results. The positive replication claims [143, 526] describe works with a certain overlap in personnel and/or a shared laboratory setup. A closer look at the timeline will be given in section 1.4.

Other types of SF trials dealt mostly with cavitation in heavy water. In 2004 Geisler et al. [161] published trials to detect fusion neutrons in coincidence with SL flashes in D_2O. The bubbles were induced with high-intensity laser shots and without any acoustic field while the water containers were either open or pressurised up to 11 bar. They recorded 58 000 SL events while running a neutron detection setup which was made for getting a significant signal if even just one in 1000 collapses would set off a neutron in any direction. They counted no events above background.

Barbaglia *et al.* [25] conducted SF experiments with a laser used for nucleating large single bubbles in D_2O and $CDCl_3$ without counting any neutrons. Experiments with laser-induced cavitation in other liquids were continued by the group [381].

Another Russian article with little reception in English-language literature describes an experiment based on the violent shock wave compression of bubbly liquid triggered by exploding a wire ring. It has been published in 2008 by Bityurin et al. [47]. The bubbly liquid with a gas volume fraction of 20 to 90 % was created by a sudden pressure drop in heavy water saturated with dissolved deuterium gas. In the moment when the pressure drop led to the birth and expansion of vapour- and deuterium gas-filled bubbles throughout the test section they discharged a capacitor bank through a wire loop placed in the liquid. The vapourisation of the wire created

[8] According to Brenner's "ballpark" estimate (see p. 11) this means that (depending on the assumed minimal bubble size) the transition into a regime is reached where the interior of the bubble cannot be treated as homogeneous any more.

[9] The Defense Advanced Research Projects Agency (DARPA) is the research arm of the U.S. Department of Defense.

[10] The BBC brought an episode of their science program "Horizon" on the question of SF in 2005 [316]. They had commissioned Putterman's UCLA group for conducting the experiment and asked a few British scientists to check the results. A transcript [317] is available on the BBC website (`http://www.bbc.co.uk/sn/tvradio/programmes/horizon/experiment_prog_summary.shtml`) and the film material itself can be found e. g. on Youtube (`https://www.youtube.com/watch?v=4hYAhPOaduM`).

a concentric pattern of shock waves propagating through the bubbly liquid. It was hoped to get as much out of the self-amplification mechanisms of the not perfectly symmetric shock wave pattern from the not wholly closed wire loop and to trigger fusion reactions in the lastly and most violently imploded bubbles in the central region. The neutron measurement method was based on capturing moderated neutrons in indium sheets placed next to the test section. Neutron capture by ^{115}In yields the excited state ^{116m1}In with a half life of 54 minutes emitting characteristic 417 keV gammas. The short half life means that quick handling is required between irradiation and gamma spectrum recording. On the other hand, it offers the advantage that the measurement time can easily cover the decomposition of most of the activated nuclei. This is unlike e. g. tritium measurements where during recording times in the range of hours only a tiny fraction of the total radioactive material decays. Bityurin et al. did in fact conclude with a positive finding, declaring that they detected a significantly raised gamma count rate in comparison with control experiments and inferring a neutron yield of 10^8 to 10^{10} $\pm 22\,\%$ per shot. This would signify a successful SF experiment, and it is a great pity that the lead was not taken up by others as it seems (verification by peers is crucial to declare something a scientific truth). In contrast to Taleyarkhan's setup where external neutrons are needed for bubble nucleation, this SF experiment involves no additional sources of radiation or radioactivity, there is no need to separate the SF neutron signature from other neutron counts which is a great advantage for data analysis and presentation.

In a paper called "Piezonuclear neutrons" Cardone et al. [73] published measurement data exhibiting a neutron signature accompanying acoustic cavitation in front of an ultrasound-emitting steel tip immersed in an aquæous solution of $FeCl_3$. They stressed that the solution contained no deuterium or other fusion fuels and that the possibility of direct H+H fusion was excluded by control experiments.[11] As it would be new physics if it was not an erroneous experiment, the group's research is much debated [74, 129].

Toriyabe et al. conducted an experiment with particle beam-induced fusion in liquid lithium in which cavitation played a role and published in 2012 [478]. They placed a little dish containing liquid lithium at around 200 °C under vacuum and pointed a 20-70 keV deuteron beam at the repeatedly cleaned free surface of the liquid. Note that this keV range corresponds to a temperature range of 2-8 $\times$ 10^8 K and means that the beam particles have similar kinetic energies as they would have in the fusion plasma of an H-bomb [517]. The beam triggered D+Li as well as D+D fusion reactions due to deuterium having been also present, dissolved in the lithium. An interesting finding was that the rate of D+D reactions was susceptible to introducing sonic cavitation in the lithium dish while the rate of D+Li reactions was not. Cavitation was introduced through a transducer mounted below the dish able to create "countless bubbles cover[ing] the surface of the liquid Li" and to raise the D+D fusion rate by 40 %. According to the authors of the paper, the temperature rise of the deuterons in the gas phase of cavitating bubbles could be a possible explanation for the finding.

[11] Worrying about $p+p$ fusion is quite unnecessary. The neutron-producing reaction channel has a tiny cross section and is extremely rare so that it represents the bottleneck reaction controlling the lifetime of stars.

1.3.5 SF setups published in patents

The cavitation fusion reactor (CFR) by Flynn [135] (described previously) seems to be the oldest patented device with the goal of achieving sonofusion. Many more patents have been filed since the 90s following the phase of vivid discussion of SL and SF in scientific literature.

In 1996 a patent was granted to Embrechts, Lahey, and Nigmatulin at Rensselaer Polytechnic Institute [126] for a cylindrical flow-through reactor. Deuterium bubbles are created in a liquid before it enters the reactor section. Transducers on the outside excite sound waves in the bubbly liquid. Special considerations are given to the conditions under which the bubbles as oscillator systems receive energy from or lose it to the sound field during their nonlinear oscillatory motions.

Around the same time Putterman and collaborators published patents [371, 372] showing a spherical flask with piezoelectric transducers glued on it and a single light-emitting bubble at the centre, implying that the SBSL setup can make fusion feasible if deuterium and tritium gas is present in the bubble. A similar setup has been patented by Pless [355] in a language aimed at the broadest generality, e. g. by not using the term bubble but speaking of matter placed at a velocity node in a resonating cavity filled with liquid of low compressibility.

For their SF experiments published in [458] Taleyarkhan et al. used a version of the acoustic resonator design developed in the 1960s by West & Howlett [505]. West and Taleyarkhan together authored two patents [464, 467] publishing the resonator design and the experimental protocol (outlined in appendix D). Taleyarkhan himself has several more patents [460–462] describing other details and potentially beneficial design modifications or extensions of the resonator.

In a patent of 1997 [274] Shui-Yin Lo describes an apparatus for collapsing seeded bubbles of oxygen and deuterium in heavy water. Bubble expansion and collapse are piston-driven. A laser is intended to ignite the oxygen-deuterium mixture in the bubbles, so that the expansion movement stabilises the bubble symmetry, whereas the chemical reaction consumes the noncondensible gas content and empties the bubble preparing it for the subsequent collapse.

A patent of A. L. Enfinger of 2004 describes a spherical transducer-driven deuterium fusion reactor which could be classified as acoustic inertial confinement fusion (AICF) reactor, however, it doesn't contain any liquid nor bubbles. The spherical inner volume is only filled with deuterium gas and transducers on the walls are to excite an acoustic standing wave pattern with a pressure antinode in the centre. The novelty is to turn the inner wall of the chamber into a mirror and to use the deuterium gas as a laser-active medium such as to create outward- and then back inward-going concentric laser avalanches timed with the acoustic motion pattern.

Janssen et al. disclosed an apparatus in 2009 [220] which is at the same time electrolysing and cavitating heavy water. The electrodes are at the same time functioning as ultrasound emitters. The proposed shape includes disc and comb structures designed to swing against each other and to create many similar compartments in the liquid volume where cavitation of D- and T-containing bubbles can occur in close proximity to the electrode surfaces.

Tomory [477] (published 2007) intends to partially break up the spherical sym-

metry of an SBSL setup and create a long tube where bubbles chained up along the central line are expanded and collapsed in waves travelling in the axial direction. The whole tube must be covered with innumerable time-controlled transducers. The patents by Agbossou & Dion [3] and by Vaxelaire [488] go in similar directions although targeting primarily at sonochemistry applications.

In the years from 2000 onwards many patents have been filed by Ross Tessien and co-workers of his company *Impulse Devices*[12]. By now, the company has grown and gone up into *Burst Laboratories*[13]. The series[14] of patents [281, 348, 349, 399, 400, 471–476] and papers[15] started around a small circle of authors which has grown in the meantime and comprises now a larger group including renowned senior scientists from the field of SL who became associated with the company. The experimental results covered within the scientific papers and the abundance of detail solutions to diverse practical problems treated within the patents suggests continuing progress in mastering more and more of the technical challenges. This singles the company and the people behind it out from the rest of the organisations and authors behind SF-related patents. The enterprise's continuing and fruitful existence is perhaps to a significant degree due to the bundling of strategic goals in the two fields of sonochemistry and sonofusion (whereby the first in contrast to the latter has already real-world applications) and the expansion into general challenges in the field of chemical industry [282]. Those patents seeming to be more attributable to SF are addressing the following issues:

- The shape of the acoustic resonator should be such that the displacement amplitude of large parts of the boundary of the liquid phase is maximised. At the same time low-displacement parts of the vessel are needed for placing supports and outlets while minimising their drag on the vibration motion. Patents on hourglass-shaped resonators with small outlets at the far ends go in that direction.

- The nesting of two vessels allows for higher sound pressure amplitudes if the outer vessel increases the static pressure and the inner vessel takes on the function of being the acoustic resonator.

- Rotation can be a way of controlling the position of bubbles and counteracting lift forces. Various versions of a thin-blade impeller system with magnetic drive and low impact on the acoustic properties of a cylindrical or spherical cavitation chamber have been disclosed in patents.

- Semi-flexible pistons capping the resonator ends allow the static pressure control of the rotating resonator without the need for guiding rods just by controlling the pressure of the gas surrounding the resonator and its rotation machinery.

[12] Impulse Devices Inc., Grass Valley, CA (USA)

[13] Burst Laboratories Inc., Grass Valley, CA (USA)

[14] only exemplary selections are given

[15] There is a publications list on the website of Burst Laboratories (`http://www.burstlabs.com/index.php/about/publications`), the latest articles at the moment of writing were [22, 151, 354, 445].

- Electron beam welding is being explored as a technique for fabricating closed chambers with small tolerances on the properties of the inner surface.

- Techniques for efficiently dissolving gasses as deuterium in liquid metals to be used as the cavitation fluid.

Thus, it seems that Ross Tessien and his co-workers at Burst Laboratories are currently developing SF-related technology with more momentum than any other group. It may also be mentioned that Impulse Devices was one of the founding members of the *Acoustic Fusion Technology Energy Consortium (AFTEC)* [255], the others being Boston University, Purdue University, the University of Mississippi in Oxford, and the University of Washington in Seattle.

There are several more patented concepts which could be listed in the context of SF, e. g. [93, 132, 322, 442] or the patent by Deeth of 2012 [109] combining the concept of the concentric laser avalanche with bubble collapse in molten glass. But further down the list the limits of physics and feasibility are getting stretched more and more. Patents are simply documented ideas and as such they do not have to be based on the prompt verifiability of experiments. The boundary conditions imposed by physics and thermodynamics become much softer in this arena of language, arguments, and standpoints than they are in the real world. Therefore, the transition between daring but sound ideas and fantasy in conflict with nature is gradual and smooth. This situation is quite aptly captured in the description of a heat energy recapture and recycling system involving a piston-driven deuterium fusion reaction chamber by G. A. Labrador [248] which is at the same time a veritable patent and an intriguing piece of artwork.

1.4 SF experiments by Taleyarkhan et al.

The claim of thermonuclear bubble fusion on a "table-top" scale published by Taleyarkhan et al. in 2002 [458] created a dispute going beyond scientific literature into popular media channels and thus was more than just a "tempest in a beaker" [409]. Fully independent attempts at reproducing the experiment with a positive fusion signature failed, and significantly, the whole story happened with the research community's *cold fusion* debacle not too far in the temporal background. As this heavily disputed experiment lies at the origin of the research question targeted herein, the experiment itself and some of the dispute deserve a more detailed look in this context chapter.

In the 1990s the investigation of SL had become a flourishing field, not the least because of the discovery of SBSL which made high-quality and time-resolved spectroscopic data available and boosted the discussion of theoretical models. The already existing thought association of SL and SF was intensified by shock-based hot spot models put forward because of their ability to explain extremely short SL pulse widths of $\approx$50 ps (or $\ll$50 ps discussed at the time) [11, 26, 28, 50, 60, 96, 187, 308, 311]. As noted before, there was a DARPA project funding the work of several academic groups on SF, the experimental work of Talyarkhan's ORNL group also belonged to it.

In 2002, an important article was published in Science [458] by R. P. Taleyarkhan, C. D. West, J. S. Cho, R. T. Lahey Jr., R. I. Nigmatulin, and R. C. Block titled "Evidence for Nuclear Emissions During Acoustic Cavitation". The authors reported on experiments conducted at the Oak Ridge National Laboratory (ORNL, TN, USA) with neutron-nucleated multi-bubble cavitation in deuterated acetone. The article was already highly disputed even in the days before its publication. The dispute was even carried to the magazine's editor-in-chief, Dr. Kennedy, who felt compelled to point out the independence of the peer review and editing process in determining all publication decisions [229, 409, 410]. An arrangement was made that Taleyarkhan et al. would cite the critical ORNL report by Shapira & Saltmarsh [415] and treat some of its criticisms already in the original article.

1.4.1 A description of Taleyarkhan's SF experiment

In order to push SL further and to enable sonofusion the group around Taleyarkhan sought to concentrate the available potential energy of a cavitation bubble onto a smaller number of gas particles forming the plasma. This meant emptying the bubbles of noncondensible gasses and the two principal measures were (a) to choose a well degassable working liquid with low accommodation coefficient[16] and (b) to rely on newly created bubbles instead of periodically oscillating ones which become stationary by accumulating[17] an equilibrium amount of noncondensible gasses. The choice of working fluid fell on acetone. Besides its low vapour pressure, it has six hydrogen atoms per molecule which can be replaced with isotopes suitable as fusion fuel, it can easily be degassed to a high degree, and like many other organic liquids it has a high cavitation strength. Chilled acetone can easily bear sound fields with tension states of several tens of bar. To give a comparison, clean degassed water cannot be brought far beyond the 1 bar limit. The measure to achieve the nucleation of fresh bubbles with precise timing at moments of highest tension was to rely on accelerator-generated neutron bursts and neutron scattering as the nucleation source. Neutron scattering creates clusters of many bubbles, initially distributed along the track of a knock-on proton or other nucleus. Bubbly regions in a liquid reduce the propagation speed of sound pressure waves. The benefit of cylindrically shaped, ellipsoidal, or spherical bubble clouds is that the delaying effect results in a converging wave of bubble implosions which can greatly increase the pressure gradients in time and space and amplify the amplitude of the incident sound pressure signal [326]. A resonator construction able to achieve the needed sound pressure amplitudes for turning tensioned acetone into a neutron-detecting bubble chamber was found in the design of West and Howlett [505]. It was a glass chamber with a cylindrical main section and flexibly mounted top and bottom pistons. A hollow

[16] Accommodation coefficients are empirical constants which can account for energy, momentum, or mass transfer balances across gas/liquid or gas/solid interfaces [375]. The accommodation coefficient in terms of energy can be defined as $(E_i - E_r)/(E_i - E_w)$, where E_i is the incoming stream from the gas side, E_r the reflected stream, and E_w is the energy stream which would occur if all reflected particles had entered (been accommodated to the) thermal equilibrium with the medium behind the interface [375]. The accommodation coefficient referred to in [254, 326, 458] is normally understood as a mass balance coefficient of condensation versus evaporation.

[17] through the process of *rectified diffusion* explained previously

cylinder of radially polarised piezoelectric ceramic glued as a belt around the glass cylinder excites the vibration motion and the internal acoustic field. Thus the main components of this SF experiment are the resonator, non-deuterated and deuterated quantities of the working fluid, cooling equipment, and the electronic equipment required to drive the piezoelectric transducer. Figure 1.4 shows a schematic of this setup.

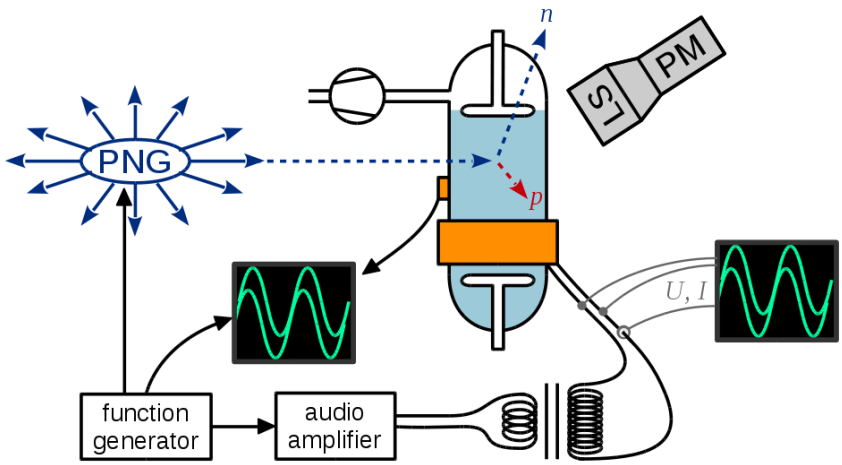


Figure 1.4 Schematic of the SF experiment setup
Besides the acoustic resonator hosting the working fluid this sketch depicts some necessary periphery equipment. A pump is needed for maintaining the low vapour pressure above the liquid surface. The equipment for driving the piezoelectric transducer comprises a function generator, an amplifier, and a transformer. The neutron source can be radioactive material or a pulsed neutron generator (PNG). A PNG has to be synchronised with the acoustic field inside the resonator so neutrons impact and can generate bubble clusters at times of maximum tension in the liquid. Two oscilloscopes are shown exemplarily for keeping track of acoustic and electric signals. A piezoelectric pill microphone glued to the outside of the resonator can pick up the sinusoidal wall motion and the high-frequency noise bursts generated by cavitation events. In the upper right, a liquid scintillator (LS) for detecting fusion neutrons is depicted. A photomultiplier (PM) tube behind the scintillator generates an electric signal which can be transformed into a recordable data stream through further electronics.

If such a setup can lead to the fusion of deuterium nuclei, then there are basically two reaction pathways available:

$$\text{(a)} \quad \mathrm{D} + \mathrm{D} \rightarrow \mathrm{T}\,(1.01\,\mathrm{MeV}) + p\,(3.03\,\mathrm{MeV}) \tag{1.1}$$

$$\text{(b)} \quad \mathrm{D} + \mathrm{D} \rightarrow {}^{3}\mathrm{He}\,(0.82\,\mathrm{MeV}) + n\,(2.45\,\mathrm{MeV}) \tag{1.2}$$

The probabilities of the two pathways are almost equal. As the reaction rates would be far too small for a chemical detection of the reaction products (helium being the only new element), the only two practically feasible ways of detecting the fusion signature rely on the **radioactivity of tritium** and the emanating **neutron radiation**.

The SF experiment has to be complemented by measurement technology able to detect one or both of these signatures. For **neutron detection** Taleyarkhan et al. installed a liquid scintillator next to the test section. Gamma rays as well as neutrons can scatter and trigger light pulses in the scintillator liquid, therefore *pulse shape discrimination (PSD)*[18] was applied to separate the gamma from the neutron

[18]The working principle of a scintillation detector is fluorescence detectable with a photomulti-

statistics.

When neutrons are relied-on at the same time for nucleating the bubble clusters and sensing the fusion reactions, then there is an inherent danger of confusing neutron counts. By the separation in time it would be easily possible to make a distinction. Whereas the neutron bursts for bubble nucleation are needed at times of maximal tension in the liquid, the fusion neutrons would eventually be created at times of maximal bubble cluster compression. With resonators driven at frequencies around 20 kHz there is a time span of 25 µs between bubble birth at the liquid's tension peak and the first collapse during the following compression phase of the acoustic cycle. This can be enough for the external neutron burst to decay (the burst duration was indicated as 12-15 µs [325, 456]). Taleyarkhan et al., however, recorded neutron statistics without interruptions because it is possible to include the external neutrons in the count and to rely fully on the comparison with control experiments. Even a small fraction of fusion neutrons can be distinguished from a many times larger amount of background neutrons if the statistics are large enough and repeated experiments remain consistent.

Taleyarkhan's team also looked for the other evidence of fusion and gathered **tritium production** data. For that purpose samples of the working fluid were investigated looking for the radioactive decays of tritium:

$$\mathrm{T} \quad \rightarrow \quad {}^{3}\mathrm{He} + e^{-} + \bar{\nu}_e \tag{1.3}$$

The corresponding half-life is 12.32 years and each decay releases 18.6 keV of energy. The beta radiation has a short stopping range in matter in the order of microns but in this particular case it poses no difficulty to detection. Samples of acetone can be directly mixed with an organic scintillation liquid and scintillation count rates can be benchmarked against control experiments. In order to increase the detection efficiency and the significance level of any signal, the acetone samples should not be too small, the error on the liquid quantities should not be too large, and the recording times should not be too short, thus there is an art to obtaining good tritium measurements.

1.4.2 The presented result data and points of critique

Nigmatulin listed the pieces of evidence which can be gathered in an SF experiment [325] as:

a) a statistically significant increase in tritium nuclei content,

b) a statistically significant number of scintillations from D-D neutrons at 2.45 MeV,

plier tube. Different excited states have different decay times and the availability of decay channels can depend on the spatial density of excited states. When neutrons scatter inside the detector they produce mostly recoil protons which in turn deposit their energy along a short track. Gamma photons scattering in the scintillation material produce a sparser distribution of excited electrons and a different composition of populated excited states. With the right choice of scintillation material it can be achieved that neutron and gamma events can be distinguished to a high degree because of their difference in the ratio between so-called delayed and prompt fluorescence.

c) an approximately equal number of D-D MeV neutrons and T nuclei produced, and

d) the generation of D-D neutrons coincident with SL flashes during bubble cloud implosions.

In their 2002 Science paper Taleyarkhan et al. reported a positive tritium signature based on samples of 1 cm^3 of acetone taken from the test section before and after cavitation runs of 7 and 12 hours. The control experiments were conducted in the form of setting the temperature to 22 °C instead of 0 °C, by cavitating natural acetone instead of deuterated acetone, and by irradiating the chamber containing deuterated or natural acetone with the external pulsed neutron generator (PNG) while having the acoustic drive (and thus cavitation) switched off. The authors inferred success in detecting SF because they measured a statistically significant increase in the tritium decay rates. In particular, the increase was only observable in the case of cavitation in chilled deuterated acetone and it was absent in any of the control experiments. The authors state the statistical significance level of the count rate increase as ranging from around 2.5 to >4.5 standard deviations. This is based on the physical nature of radioactive decay processes which can be described by a Poisson distribution. From the data given in figure 3 of [458] and the 7 to 13 % coefficient of variance[19] it can be inferred from the count rate that measurements were carried out over time periods of 4 to 5 minutes. Obviously, a few minutes of added measurement time could have greatly improved the measurement precision in terms of statistics. For the lowest reported count rates of 15 cpm (i. e. counts per minute) going from 5 to 20 minutes would have reduced the standard deviation by half. Imagining how the statistical error on decay rate measurements can easily be reduced quite far in such a case makes it also clear that a context discussion of other possible sources of error, like e. g. the error on measuring one cubic centimetre of acetone sample size (see [390]), can become very meaningful. In that sense one could say that the presentation of some key arguments in the Science article by Taleyarkhan et al. could have been better. In fact, subsequent publications of the author group on the series of SF experiments [325, 395, 455, 456, 459, 465] provide significantly extended and improved data and discussions.

Also in terms of evidence (b), i. e. neutron detection with a scintillator and PSD, the report reached a positive conclusion. At count rates of about 500 neutron counts per second data was gathered over periods of 100 and 300 s. When hitting protons neutrons can deposit all their kinetic energy of 2.45 MeV or less in the scintillator. The case for a successful proof of SF was made because significant count rate differences were listed for two types of comparisons: cavitation on versus cavitation off on the one hand and pulse heights corresponding to deposited energies above 2.45 MeV (the D-D fusion energy level) versus pulse heights below that threshold on the other hand. The SF-related increase of 4 % in the number of counts below the energy threshold was associated with a significance level of greater than ten standard deviations. No time windowing seems to have taken place, i. e. detector counts recorded after bubble implosion were collected in the same statistics as counts recorded in

[19] The coefficient of variance CV is the ratio between the standard deviation and the absolute signal $CV = \sigma/N$ and equates to $\sqrt{N}/N$ if Poisson statistics are applicable.

close time proximity with the PNG pulse. As a reader one wonders also why the recording times have not been enlarged to yield yet clearer signatures. In [455] it is already mentioned that the PNG's neutron output underlied a variation of $\sim 0.2\,\%$ from measurement to measurement. As can be inferred from a later published discussion of the experimental data [325, 465], the authors stress that long recording times are prone to pick up the drift in conditions as signal whereas short periods and often switching from cavitation to the cavitation-off control experiment yields more robust and reliable data. They underlined the repeatability of the effect while other lab conditions were kept the same.

In terms of evidence (c) the observed neutron count was described in the original paper [458] as more than one order of magnitude too small to fit the tritium production data, which was in the original article attributed to estimation errors of the neutron detection efficiency and the possibility of an inhomogeneous distribution of the tritium content in the test section. The mismatch between presented neutron and tritium data has received particular critique in [178, 395] (and has been mentioned also in [415]). Together with responses of Taleyarkhan and his co-authors in i. a. [325, 455, 456, 465] this represents a dispute on the technicalities of nuclear instrumentation, a dispute stretching over years concerning one and the same old dataset. Questions of nuclear instrumentation can sometimes be nontrivial, but nevertheless such questions are settled regularly in many nuclear labs throughout the world. As rightly pointed out by Taleyarkhan et al. [395][20], the calibration of complex detector and instrumentation setups with radiation sources of known intensity is a common and efficient technique and always preferable to theoretically calculating a detection system response based on partially calibration and partially literature data. This facet of the SF controversy alone bears witness of a need to suggest improved experimental setups, so that new SF trial reports (to be planned and funded) can promise to add more substance to the dispute.

The last piece of evidence, the neutron-SL coincidences, were evaluated by collecting the neutron counts in histogram bins according to their time offset with respect to SL flashes. The histogram bins represented time windows of $2\,\mu s$. Also in this case the authors presented a visible fusion signature (more coincidence counts with cavitation than without, no difference in the case of natural acetone) and deemed it statistically significant as the difference was (in repeated measurements) larger than one standard deviation. The coincidence data received critique from Saltmarsh & Shapira [395] because the detection system stayed active also during PNG firing times whereby it could become prone to elevated numbers of random coincidences. A different critique was raised by Putterman et al. [367] who disapproved of the large time windows. In their opinion, "true coincidences" would be much more efficiently filtered with time windows in the nanosecond range. Indeed, SL flashes are, as was mentioned above, very short events, mostly staying below the nanosecond scale, and 2.45 MeV fusion neutrons travelling at $\sim 7\,\%$ of the speed of light cover about 2 cm per nanosecond. As a related matter of fact, detecting positron annihilation events through gamma coincidences in nanosecond time windows is a common experiment for physics students. However, it has to be kept

[20] in the second part, the response part of the technical note

in mind that γ-γ coincidence measurements require a pair of identical scintillation detectors which can have a fast photon emission decay characteristic. By contrast, for SL-n coincidence detection an asymmetric setup is needed with a bare PMT on one side and a scintillator with PSD requiring slower decay times on the other side. Another important aspect is that when a bubble cluster implodes it means a narrow sequence of bubble implosions from the outside towards the centre of the cluster. Trying to resolve single bubble implosion events instead of summing over the whole cluster implosion means that challenges arising from signal sparsity and detector dead times will increase significantly.

1.4.3 Sketching the unfolding dispute

What prevented the story of SF from proceeding along a more regular pattern was foremost that neither a prompt independent verification was achieved by anybody, nor did a falsification appear as in the form of a convincing alternative explanation of reported positive measurement data. The view that it can't be ruled out by principle that SL plasma might under particular conditions become hot enough for fusion was and still is held by many researchers. A long list of references [97, 153, 178, 251, 254, 273, 318, 319, 324–326, 395, 414, 453, 454, 456, 457, 459, 463, 465, 466][21] shows that the initial article triggered a whole series of criticising comments, answers, follow-ups, and new critiques. The discussion mirrored by that list mainly deals with technical issues of how the neutron and tritium data were gathered (as outlined in the preceding section), i. e. how the detectors were set up, what these setups would be able to see, and how the data were postprocessed and presented. Besides articles and letters in scientific journals, the debate was accompanied by many magazine and online articles, e. g. [6, 32, 80, 83, 175, 183, 255, 268, 346, 397, 409, 490], and has already been described by Seife in a book [410]. The above-average online and news coverage has of course more reasons than it just being an interesting experiment which could not promptly be replicated or otherwise explained. It has to do with the history of the decades-old quest for controlled thermonuclear fusion [410] on the one hand, and on the other hand with the sometimes carelessly added, sometimes purposefully invoked image of fusion as an advanced and climate-neutral technology for electricity production with potentially disruptive character. Another aspect is that the work and publications of Taleyarkhan et al. have also received criticism and been involved in arguments in the context of research standards, ethics, and politics (see e. g. [79, 82, 183, 241, 383–385, 410, 485]). This latter aspect will however not be addressed in the current discussion.[22] A comprehensive collection of articles, material, and links on the "Bubblegate" story has been put together by Steven Krivit et al. of the New Energy Times[23] and can be found on the magazine's website.

[21] Replication trials are excluded here but discussed below.

[22] Firstly, a scientific publication is not the appropriate place to add anything to this dispute. Secondly, it is assumed that a justified interest in the question of SF is best helped by advancing the scientific part of the debate by pointing out some underexposed but perhaps crucial technical issues of past experiments and their reproducibility. It can also be noted that the ORNL SF reports are not within the scope of these criticisms.

[23] link: `www.newenergytimes.com`, publisher: Steven B. Krivit

Replication trials

Claims, hypotheses, and bold new experiments are essential to the progress of natural sciences, but ultimately they are worth nothing as long as their condition of independent verifiability is unclear. For experimental research this means that the acceptance of scientific news hinges on the reproducibility by independent researchers or teams, i. e. the generality of reproducibility. For the SF experiments reported by Taleyarkhan et al. the status of reproducibility is very unclear. There have been unsuccessful replication trials of which it can be said that the followed protocol exhibited strong deviations in crucial aspects and there have been claimed successful replications, however none without any overlap in lab equipment, personnel, or person relationship network.

Shapira & Saltmarsh (ORNL): The SF experiment reported by D. Shapira & M. J. Saltmarsh [414–416] was often cited as an early failed replication. It was actually a collaborative project with Taleyarkhan and his team running the experimental machinery in their ORNL lab and Shapira & Saltmarsh from the physics department of ORNL recording neutron data with their detectors. The authors compared data collected during a one-hour cavitation run with data from a subsequent one-hour run with cavitation switched off and concluded that there was no significant signal observed which would have confirmed SF. However, as pointed out by Taleyarkhan et al. [455, 465], the SF experiment under the guidance of Shapira and Saltmarsh was reduced in scope and modified in its protocol. The scope was reduced by not collecting tritium data taking away the possibility of a quantitative comparison of the two types of fusion signature. As the most decisive deviation of protocol the sequence of cavitation and control run periods followed a different pattern, just two single long runs, which made the experiment susceptible to drifting experimental conditions. What may have raised the level of confusion of many followers of the SF debate even more is that Taleyarkhan et al. postprocessed and presented the raw data gathered by Shapira & Saltmarsh (to which they had been given access as collaborating colleagues) in a different way and drew the conclusion that it did indeed exhibit a positive SF signature.

At Purdue University in West Lafayette, Indiana, two groups tried to stage replications of the ORNL SF experiments. Both were in contact with Taleyarkhan et al. already during the development and construction phase of their experiment setups. In summer 2003 Taleyarkhan himself moved from ORNL to Purdue where he had accepted a chaired professor position. Of course this also meant closer contact and more occasions for knowledge transfer. Thus, the two Purdue groups can at least not be considered totally independent.

Xu, Ravenkar & Butt (Purdue): The first Purdue group led by Yiban Xu reported positive results. Xu authored a paper together with Adam Butt which appeared in Nuclear Engineering and Design (NED) in 2005 [526] and a conference article [524] presented in the same year at NURETH-11[24] with Butt and Shripad T. Ravenkar. They reported positive tritium and neutron signatures. The authors acknowledged "advice, guidance and assistance" from Taleyarkhan and Cho, two

[24]The 11$^{\text{th}}$ international topical meeting on nuclear reactor thermal hydraulics (NURETH-11) was held in Avignon, France.

of the authors of the original Science paper [458]. As members of Purdue's School of Nuclear Engineering, headed by L. Tsoukalas at the time, the authors shared laboratory infrastructure and regular meetings[25] with other Purdue faculty and staff.

The independence of Xu and Butt became an issue of heated debate. Around the time of these SF experiments, Adam Butt became one of Taleyarkhan's research group members. In 2005 he submitted a master thesis titled "Acoustic Inertial Confinement Fusion: Characterization of Reaction Chamber" [66] of which Taleyarkhan was the advisor. That Taleyarkhan et al. (including Y. Xu) cited the NED paper [526] of Xu and Butt as independent confirmation[26] [459] of what Taleyarkhan et al. (without Xu) had reported in the 2002 Science paper [458] and [325, 326, 456] later became the origin of great criticism. It led to allegations of research misconduct, investigations by several committees, and finally had adverse consequences for Taleyarkhan's professional career [241]. Nevertheless, Taleyarkhan and his co-authors upheld the view that the SF experiment conducted by Xu & Butt at Purdue was independent of the ORNL experiments. Such an opinion can e. g. be found in an affidavit written by C. West for an investigation committee [506][27]. In the manuscript West stresses the notion of independent confirmation based on the equivalence of the observations and the differences in personnel, experiment configuration, location (laboratory, state, institution), the method of bubble nucleation, test cell, radiation detection systems (which they calibrated), all of which would have allowed Xu & Butt to "obtain their own data" and derive their own observations.

Tsoukalas et al. (Purdue): The second Purdue group published an article reporting negative results in 2006 [480] based on an experiment campaign[28] including the second half of 2003 [53]. The authors were L. Tsoukalas, F. Clikeman, M. Bertodano, T. Jevremovic, J. Walter, A. Bougaev, and E. Merritt. In a similar way as the other Purdue group, they were in contact with Taleyarkhan et al.. Tsoukalas et al. conducted a multitude of seven-hour cavitation runs in chilled deuterated acetone and a few control runs with normal acetone or irradiation only. They also presented a vertical pressure profile taken with a hydrophone exhibiting a unimodal shape with a central antinode and two nodes at the ends (at depths of 0 and 80 mm). In their article they described the protocol they used as scanning the frequency range between 16 and 22 kHz and then targeting the fundamental resonance mode. This

[25] according to [525], an affidavit by Xu available under `http://newenergytimes.com/v2/bubblegate/Aff/Xu-Yiban-Jan31-2008.pdf`

[26] In [459] the authors stated that "these observations have now been independently confirmed." Given that the SF claims by Taleyarkhan et al. had received heavy criticism from the beginning (i. e. four years before) and that no independent confirmation had been reported since, it is obvious that this sentence in the opening section is instrumental in supporting their case and not a mere side note. Adding no further context to this sentence was surely not a good decision.

[27] The document can be found in the resource part of the "Bubblegate" section of the "New Energy Times" website edited by Steven Krivit under the link `http://newenergytimes.com/v2/bubblegate/Aff/West.pdf`

[28] According to Steven Krivit of New Energy Times the group started its efforts in early 2002 and managed to produce good cavitation rates in stable resonators only after a visit of some members to ORNL for receiving knowledge transfer from Taleyarkhan and his group. The report [240] is online at `http://news.newenergytimes.net/2014/01/20/federal-investigations-reveal-academic-backstabbing-at-purdue-university-part-3/`.

paper [480] increased the confusion around SF because a draft version [53][29] of it was circulated which ended with a positive conclusion based on a tritium result graph containing different results.

Camara et al. (UCLA): Within the framework of the same line of DARPA grants which also supported SF experiments by Taleyarkhan et al., it was agreed that within the research groups collaborating for that grant an independent replication of Taleyarkhan's acetone-based multibubble SF experiment should be staged. This experiment with the aim of being a "carbon copy" was conducted at UCLA by C. G. Camara, S. D. Hopkins, K. S. Suslick, and S. J. Putterman and published in 2007 [67]. It yielded a negative result. Camara et al. stated that "fusion [...] is zero within our experimental accuracy" which would correspond to an upper bound of 200 neutrons per second. However, a close look at figure 1 of the paper reveals a decisive difference in detail, namely, that the test section was not filled all the way to the top piston with the working liquid. In the figure caption the authors mentioned as justification that it "resulted in more stable acoustic modes and avoided excessive breakage" and stated that it was the setup applied in "most runs" covered by the report. It is clear that the introduction of a gap between the liquid and the upper piston corresponds to a severe change in the acoustic properties of the resonator which is why the UCLA experiment cannot be called a "carbon copy" at all. A possible other deviation from the ORNL SF setup is the introduction of air into the previously degassed acetone in order to raise the SL intensity. Camara et al. wrote that "at about 20 torr of air, the observed SL signal was similar to that reported by Taleyarkhan et al." but in the article they left unclear how much of their experiment campaign was affected by this additional deviation in protocol. The SF replication trials described in [67] covered also a replication trial by Putterman and his team conducted for a BBC documentary [316, 317].

R. Tessien (Impulse Devices Inc.): Ross Tessien and his team of Impulse Devices Inc.[30] have tried to replicate the SF experiment of Taleyarkhan et al.. There are no publications and little[31] is known about the trials. The case of Tessien allows the assumption that there have been perhaps some more unpublished replication trials.

Forringer, Robbins & Martin (LeTourneau University, Longview, TX): In May 2006 physics professor Edward R. Forringer and two of his students, David Robbins and Jonathan Martin, went to visit the meta-stable fluids research lab at Purdue where they were welcomed by R. Taleyarkhan and Y. Xu in order to perform SF experiments with the lab's equipment but using their own track detectors for neutron counting. Over three days they performed SF experiments with a mixture of deuterated acetone, deuterated benzine (petroleum ether), tetrachloro-ethene, and uranyl nitrate. Forringer et al. drew the positive conclusion that they had indeed observed fusion and presented their work as a conference paper [143]. Some more

[29] It can be found under `http://newenergytimes.com/v2/bubblegate/2004/2004-TsoukalasL-TritiumEvidence-Draft.pdf`

[30] today Burst Laboratories Inc., Grass Valley, CA (USA)

[31] Tessien's replication trials are mentioned in [81, 200, 485]. According to Chang [81] Tessien "abandoned using Dr. Taleyarkhan's approach", and Vance [485] cites him saying "It's a nightmare to run it, and it breaks".

details can be found in [142]. These texts state that the experiments were conducted after explanations and with assistance from Taleyarkhan and Xu but under own and independent guidance. In particular, the experiment campaign included the use of a liquid scintillation detector, its calibration, and the recording of a neutron spectrum from a ^{252}Cf source in order to exclude californium as a source of neutrons which might be confounded with the fusion signature. This part of the experiment was added in order to address suspicions[32] raised a few weeks earlier by B. Naranjo and his colleagues of the UCLA group [319]. To summarise, the work of Forringer et al. can be described as a repetition of the SF experiment with a modified working liquid but complete overlap in resonator and driving equipment.

W. Bugg (University of Tennessee): William Bugg was another visitor[33] to Taleyarkhan's lab with a strong interest in witnessing a full cycle of SF experiments. Taleyarkhan conducted for him an experiment of reduced scope within a one-day session allowing Bugg to gather neutron data from two runs with deuterated and normal liquid with track detectors. The employed working liquid was a "benzene-acetone mixture with a dissolved uranium salt" for bubble nucleation. The work was not published and the only publicly available documentation is hosted on the New Energy Times webpage [63, 64]. According to these descriptions, Bugg witnessed a demonstration, oversaw the random choice and placement of track detectors, the cavitation in deuterated and normal working liquids, the etching of the detectors, and subsequently he analysed the detector traces himself. However, one can not classify this as independent confirmatory experiments.

Saglime, Danon, Lahey et al. (RPI): Soon after the publication by Taleyarkhan et al. in 2002 [458] a team at RPI under the initiative of Professors Yaron Danon and Richard T. Lahey Jr. started an attempt to repeat the ORNL sonofusion experiment independently at the Gaerttner Laboratory, the linear accelerator lab of Rensselaer Polytechnic Institute (RPI, Troy, NY). The team included Frank J. Saglime III, Robert C. Block, Yaron Danon, Richard T. Lahey Jr., Francisco Moraga, Silvina Cancelos, Jason Brodsky, Justin Walraven, and William Schlichting. They started out with resonators assembled mainly from glass parts received from Taleyarkhan's Purdue team and continued building their own resonators of varying shapes. Their SF experiments targeted both neutron and tritium detection and encompassed several experimental campaigns employing different resonators. For cavitation nucleation a Pu-Be neutron source was available at the Gaerttner Lab, but there was also the large, accelerator-driven pulsed neutron source at the heart of the lab. The SF experiments for the neutron beam-nucleated setups were located in the flight station along the neutron beam lines closest to the neutron source which was at a distance of 25 m. With the Pu-Be-nucleated setups Saglime et al. were seeking the neutron and tritium signatures of SF. The beam-nucleated setup was used only for the purpose of gathering fusion neutron data because of the high costs

[32] Naranjo's paper does not literally raise an allegation of fraud, but to the informed reader the implication is straightforward, as any (isotope-based) fake neutron source or it's shielding would have to change place during demonstrations when switching between SF and control experiments.

[33] By his vita William Bugg can be described as experienced visitor with particular expertise in particle and radiation detection. He is a senior experimental physicist in the field of high energy physics and long-term associate of Standford's SLAC facility.

related to hour-long accelerator beam operation. In their reports [250, 252, 390] the RPI team concluded on not having observed any statistically significant traces of SF. As notable differences between the ORNL setup and their's they listed the pulsed neutron source, the acoustic resonators, and the electrical equipment.

However, there is one part of the data analysis in which the negative conclusion was not drawn in a straightforward manner. It is a tritium level measurement plot from three long-term runs with the Pu-Be source and deuterated acetone: one 10-hour run, one 24-hour run, one 24-hour run without cavitation. The three corresponding samples were compared in their radioactivity with a fourth fresh sample of working liquid. This plot[34] indicates an elevated level of accumulated tritium with respect to the two control samples which would be statistically significant if judged only by Poisson statistics. For the 24-hour run the significance level would be about six standard deviations. Furthermore, the time-dependence of the radiation build-up looks to be not far from linear. However, the authors of [250, 390] added a second analysis and took into account an additional existing source of error, namely the precision error on taking small samples of 1 ml of acetone by means of a pipette. In the second analysis the increase in tritium was not statistically significant any more.

Later, Saglime et al. repeated the same tritium build-up experiment, this time verifying the mass of the extracted liquid samples with a precision balance. The result of the data analysis then yielded a negative result with more clarity [393].[35] More details on the SF-related work at RPI can be found below in section 1.5 (p. 33) and in appendix I.

1.4.4 Summarising the status of the condition of verification and replication of the SF experiment according to Taleyarkhan et al.

One can say that given the history of publications and replication trials outlined above, the situation allows some room for interpretation. A sceptical reading of the facts could be summarised as: the only positive published results come either from experiments conducted directly in Taleyarkhan's laboratories and with his equipment (Forringer et al.) or from a group within the same department of the same university with significant overlaps and relationships in persons and lab infrastructure (Xu et al.). Several other replication trials have yielded nil results, including completely independent ones. Thus, one cannot say that there has been a completely independent successful replication of sonofusion based on acoustic cavitation in organic liquids since Taleyarkhan et al..

However, the basis of facts also allows an extended interpretation: there have been too few replication trials and some of these few are tainted by surrounding issues or even severe flaws such as the flawed resonator setup at UCLA. Therefore, the current status holds room for ambiguity and is unsatisfactory, so that further replication trials of high quality remain desirable, in particular because the physics

[34] figure 67 on page 92 in [390] or figure 37 on page 51 in [250]

[35] This data from the later part of the SF experiment campaign of Saglime et al. was never published.

of SF is interesting (i. e. extreme energy concentration and SF as quantitative probe) while the experimental setup itself is rather simple and relatively inexpensive.

1.5 Research on sonofusion and cavitation resonators by Lahey, Block, Danon, Saglime, Cancelos et al. at RPI

Two of Taleyarkhan's co-authors of [458], Richard T. Lahey Jr., and Robert C. Block, as members of the Rensselaer Polytechnic Institute (RPI, Troy, NY, USA), saw the chance to take advantage of RPI's nuclear physics infrastructure for staging re-trials of the SF experiment with a high degree of independence from the labs at ORNL and Purdue. As previously discussed, the Gaerttner Laboratory at RPI hosts a linear accelerator, the *RPI Linac*. The Linac offers a special working mode as a pulsed neutron source when placing the right target materials in the electron beam suitable for maximising the output of so-called photoneutrons[36]. The possibility to move an experimental setup into the beamline of one of the available outgoing neutron flight tubes is one of the main benefits. As in any other nuclear lab small mobile radiation sources are also available for instrumentation setup and calibration purposes, among them a Pu-Be neutron source. The RPI work before 2007 can be divided into two blocks. In a first phase replications of the SF experiment were directly attempted by Saglime et al.. As these efforts brought about no positive results, they were followed by a second phase of investigating and designing acoustic resonator setups. At that stage the goals were to make the performance of the old-style resonators more reliable and to come up with a new more robust design which can be assembled from machined parts. As part of this introductory chapter, the following paragraphs will serve as a short outline of the prior RPI-based SF research activity. Complementary and more detailed documentation materials are added in appendix I.

SF replication trials

The SF research at RPI started out with straightforward attempts at replicating the ORNL SF experiments. The RPI team even received used resonators and resonator parts from Taleyarkhan's new Purdue team. Going out from there, a collection of own resonators was manufactured (see the list in appendix I.4). With at least two of them SF trials were conducted by Saglime et al. [250, 252, 390]. Concerning neutrons, these SF trials followed the same pattern as described in [325, 458]. Cavitation was repeatedly switched on and off and the data from many on-off recording period pairs were collected in cumulative count statistics which made the experiment robust against long-term drift of experimental conditions. Control experiments with natural acetone were conducted for comparison against the data from deuterated acetone. The RPI team could not find any significant neutron signature of SF.

As running the RPI Linac is much more costly than using a small-scale PNG (as done at ORNL), SF runs for collecting tritium data were conducted with the Pu-Be

[36] Appendix G.3 explains in basic terms what photoneutrons are.

neutron source for cavitation nucleation instead. Indeed, there seems to be a clearly visible positive SF signature in the data published in [390]. But this signature was determined as insignificant after having taken into account the possible error on the sizes of the liquid samples. The (unpublished) repetition of the experiment with an added step of measuring the acetone sample sizes on a precision balance yielded negative results with more clarity [393].

Development of resonator design

The RPI team built a series of resonators of which several exemplars can be seen in figure I.5 (p. 292) in the appendix. These resonators can be classified into variations of the old West-Howlett style and the new design with H-shaped liquid domain cross sections introduced by S. Cancelos [69]. Schematics can be seen in figure 1.5.

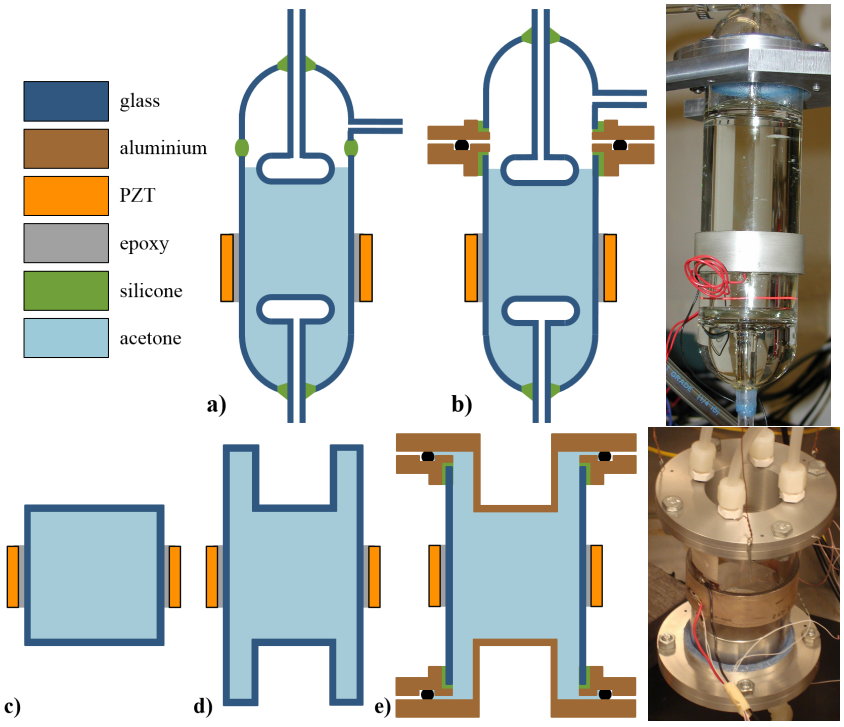


Figure 1.5 Schematics of resonator designs examined at RPI
The RPI team examined acoustic resonators of the West-Howlett design [505] (upper row) and an alternative design with H-shaped fluid domain cross sections developed by Cancelos [69]. With only one horizontal nozzle for degassing in the upper head and the latter being glued by silicone to the main section, refilling the resonator required cutting up and remaking of the silicone bead. The introduction of the bolt-clamped aluminium flange had the purpose to alleviate this situation. The new design with an H-shaped cross section of the liquid domain was constructed so that one half wavelength of the standing acoustic wave would fit in horizontally as well as vertically into the main segment of the liquid domain. The H-shape design (d+e) is a slight complication of the trivial cylinder geometry (c) with the inherent problem of stress concentration in the edges. The extensions of the cylindrical side wall above and below are to minimise deformation levels and thus stress, fatigue, and damping at and around the corners.

Going through the old-style resonators in figure I.5 one can see various variations of assembly techniques, piston and chamber fixations, outlet seals and connections, and so on. This bears witness of the challenges with respect to assembly integrity, part positioning, vacuum tightness, and handling procedures. It can be seen that

different types and positions of piezoelectric transducers have been used for investigating different vibration modes. On top comes the challenge of drifting working conditions, in particular a drifting resonance frequency which has been attributed to temperature effects and the steady shrinking of the acetone content in the test section due to evaporation. Whereas in the ORNL experiments for [458] the experimenters had to keep the resonator in tune manually, a LabVIEW® code was developed at RPI to periodically adjust the driving frequency to the resonator's drifting resonance frequency. There are basically two useful measures for the quality of the resonator and the momentary working point: the sound pressure gain, i. e. the achieved sound pressure amplitude divided by the driving voltage, and the rate of bubble bursts per second. Enlarging the amplitude of the standing acoustic wave enlarges both the time window and the size of the area where the tension in the liquid breaches the threshold for radiation-induced bubble nucleation, this is how the two measures are coupled. However, as the sound field-perturbing and cavitation-inducing hydrophone can only be used during characterisation experiments at low amplitudes, the bubble burst rate becomes the only relevant performance measure during SF runs.

If resonator performance changes from resonator to resonator and for one single resonator from hour to hour it has important consequences:

- It is detrimental for measurement campaigns. If evaporation of the working liquid limits the time span of single experiments, and if refilling requires chamber opening and re-sealing which entails even further performance jumps, then the negative impact is even more severe.
- If "cavitation-on" neutron statistics are to be compared with "cavitation-off" statistics, statistical significance can much faster be reached if the influence of the working condition drift is systematically eliminated by cumulating data from many repeated pairs of cavitation-on and cavitation-off runs.
- Even if positive SF signatures were resulting from single experiments, how useful is that? Given the in-house reproducibility limits experienced by one team of experimenters, it must be extrapolated that the reproducibility challenges existing for any independent team involving different glassblowers and people assembling the parts and conducting the experiments will be much, much higher, particularly if all are newcomers to this type of experiment which would be the precondition for independent verifications.

Therefore, the efforts of the RPI team began to focus on better reproducible and more robust resonator design.

New resonator designs by S. Cancelos

An acoustic resonator design was developed at RPI by S. Cancelos. It has an H-shaped fluid domain cross section and is depicted in the lower row of figure 1.5. Four versions were built: two larger ones for cavitation in water at around 14 kHz (for use in biophysics research) and two smaller ones for cavitation in acetone at around 20 kHz. Of each pair one test section was made wholly of glass while the other one

consisted of a glass cylinder segment with glued-on aluminium flanges and bolted end caps. Cancelos used the larger all-glass resonator for validating FEM analyses and the aluminium flange version for cavitation runs with water. The resonator achieved the required cavitation rate with a driving voltage of 40 V while dissipating 6 W of electric power.

This resonator design had several advantages. Both the all-glass and the flange-equipped version can be assembled from segments of industrially produced glass tubes and plates and precision-machined metal parts. There is no free liquid surface inside the resonator volume any more and the connecting tubes can stay filled with acetone during operation. Thus, an important source of drift in acoustic properties for previous designs was eliminated. Lastly, the nozzles can be placed at the highest and lowest point of the resonator's inner volume where they allow remote chamber filling without residual gas bubbles. At the same time these locations minimise the acoustic coupling to the connected tubing.

As small size variant for cavitation in acetone the all-glass version was built first. It exhibited a very high Q-factor [393] but early on during characterisation measurements and cavitation trials the glass cracked due to fatigue around the top rim where the wall curvature was large and deformation levels, too. A version with aluminium flanges and end caps had been built but was not used.

1.6 Research on sonofusion and cavitation resonators by Lahey, Malouin, Stokmaier et al. at RPI & KIT

In 2007 a collaboration began between RPI and KIT in which this author travelled to RPI to take part in renewed experimental efforts together with Bernard A. Malouin, and supervised by Richard T. Lahey Jr., with the goal of conducting new SF experiments. It was intended to start with the small version of Cancelos' resonator design with the aluminium end caps as it had reliably enabled sufficient cavitation in water. An exemplar had already been built previously, it is the resonator labelled Nº 8 according to the listing in appendix I.4. However, it turned out that the resonator exhibited such a high damping rate that it was rendered unsuitable for cavitation runs within the existing infrastructure. It required driving voltage amplitudes near 1 kV in order to yield decent cavitation rates in the presence of the Pu-Be source. The system of electronics and transducer turned nonlinear at that point which became visible through large anharmonicities in oscilloscope traces. At the same time the energy dissipation became too large for being handled by the cooling system based on cold air fanned towards the resonator, it heated up and by consequence the cavitation bubble bursts turned into mere boiling of the liquid. The resonator's temperature and boiling would decay slowly after switching off the acoustic drive. The measured electrical power input at elevated driving voltages (but still far below the cavitation threshold) is plotted in figure 1.6. It is no surprise then that this resonator also exhibits a low Q-factor. Q values inferred from recorded characterisation data range from 15 to 60 (see tables O.2 & O.3, pages 373 & 378).

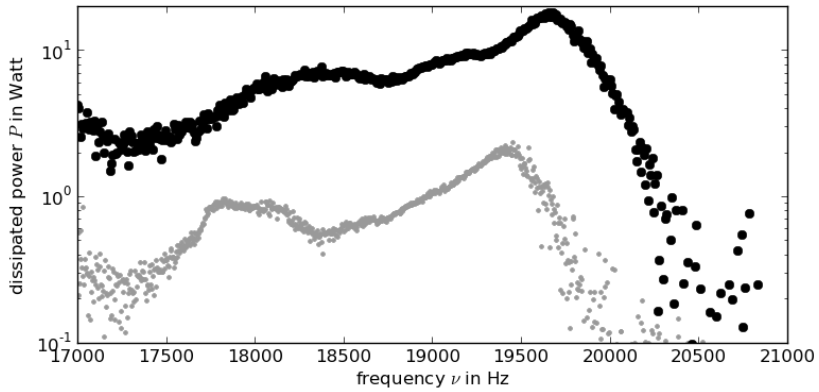


Figure 1.6 Resonator Nº 8: power dissipation at elevated driving voltage.
This plot shows the dissipated power over the frequency for two settings far below the cavitation threshold. While the peak driving voltage increased by a factor of 2.1 from the grey to the black dataset corresponding to driving voltage amplitude peaks of 159 and 333 V, respectively, the dissipated power grew by a factor of 7.7 (from 2.3 to 18.1 W). Another doubling of the driving voltage would be needed to breach the cavitation threshold of −7 bar from there, so extrapolation would imply ∼ 150 W heat generation. This means that during cavitation runs this resonator produces a heat load of the same order of magnitude as a light bulb.

Further advancing the project with simulation-aided design concept development

At that point it was decided that the best way to progress the SF project would be to try to answer the question of what made one resonator good and the other one perform badly, and to pursue that goal with a combined effort of improved resonator characterisation measurements conducted at the Linac lab at RPI together with improved FEM simulations which could be benchmarked against and calibrated with the collected measurement data.

The following experimental campaign included a thorough investigation of the test section exhibiting the most symmetric shape among the existing resonators of the conventional design. The one with the most evenly shaped glass assembly parts and silicone glue connections was chosen because one could expect the closest match between the 3D reality and a 2D FE model. As a second resonator to be examined, the underperforming new flange-equipped resonator was selected. The data collected came from a hydrophone positioned along the central axis in the liquid inside the resonator, from pill microphones fixed on the outer glass surface, from current and voltage signals picked up on the transducer supply lines, and from a needle with magnet-coil transducer able to scan the displacement of the glass surface. Providing the displacement signal was one innovation, the other principal improvement consisted in the establishment of complete 2D maps of amplitude and phase signals over the axes frequency and vertical position for both the hydrophone and the displacement data. The difference in damping ratios and mode shapes between the two resonator designs became clearly evident, and a data library was compiled which would be very suitable for benchmarking FEM simulations. The results of that measurement campaign are added as context to this thesis in appendix O.

A first 2D-axis-symmetric FEM model was created for the benchmarking case

of the opened conventional resonator with the hydrophone in place. The FEM software suite ANSYS® (Classic/APDL) was chosen because it offered at the same time comfortable parametrised geometry generation and the inclusion of the physics of the piezoelectric ceramic forming the transducer. With repeated simulations and a varying hydrophone position an equivalent pressure map was compiled to be compared with its experimental twin, and based on dedicated postprocessing the same comparison was enabled for the displacement maps. These comparisons showed that the model was indeed able to reproduce the feature-rich pressure and displacement maps qualitatively, in that the main mode shapes appeared and were in the right sequence, although the situation was still far from a perfect match. That a perfect match can in fact not be expected in this case becomes clear when taking into account the insights which were gained during subsequent simulations.

The next investigated geometry was that of the same resonator in its closed setup corresponding to SF trials with the upper piston and the top head in place instead of the hydrophone. A sensitivity study of 91 simulations varying 15 geometry parameters in 7 steps revealed that there are many parameters which have a great influence on the pressure map and that for some parameters fractions of millimetres are decisive. How the dimension deviations impact on the peak sound pressure is shown in figure 1.7. This means that no tight match can be expected for any FEM model outputs unless a tight match is made available on the side of the geometry base, either by e. g. going out from a 3D tomography image of a resonator represented by a 3D FE mesh or else by transitioning to resonators manufactured with tightly controlled tolerances. The results of the FEM model analyses of SF resonators are compiled in appendix Q. The corresponding theoretical basis is outlined in appendix chapters J (transducers) and K (finite element method).

The refined experimental resonator characterisations in conjunction with the thorough examination of the FEM model yielded two crucial insights, namely that unsystematic design trials will turn resonator construction into a game of luck, and secondly, that any different design ideas can only be usefully compared after thoroughly parameter-tuning each one.

Firstly, on the challenge of constructing a good resonator one can conclude that if one experiments unsystematically with assembly part shapes and proportions (e. g. piston shapes and positions), it is quite improbable to achieve a resonator with outstanding performance where all masses, stiffnesses, and dimensions fit perfectly well to achieve nearly the highest possible sound pressure gain and Q-factor. With a good resonator there will be reserves available: as thought experiment we can imagine the acetone level dropping below the ideal range due to evaporation; within limits (given by the electronics and the cooling system) it will always be possible to compensate for a loss in pressure gain by cranking up the driving voltage for restoring a relevant performance measure as the bubble burst rate. But as soon as resonators are assembled in a way so that masses, stiffnesses, and dimensions do not fit very well, it may be the case that even elevated driving voltages will not be able to produce a desirable sound field. It has to be kept in mind that the sound field not only has to fulfil a requirement in terms of peak amplitude because there are also requirements in terms of the sound field topology which should enable the spherical collapse of bubble clusters and prevent cavitation near the walls.

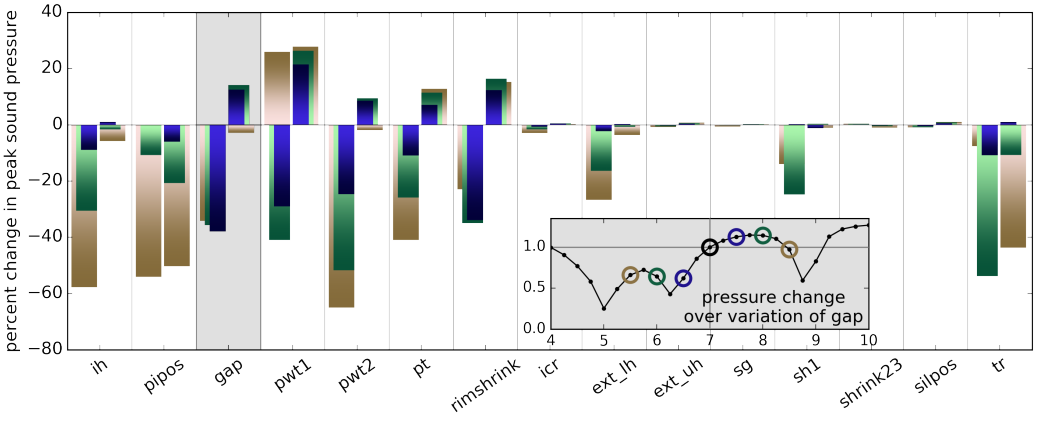


Figure 1.7 Sensitivity study of the West-Howlett resonator
The results of a senstivity study are visualised where 15 parameters were each increased and decreased by three steps. The step sizes for the parameters reflected ad hoc assumptions of the manufacturing and assembly tolerances. The investigated response is the peak sound pressure amplitude found along the central axis in the main liquid volume bwetween upper and lower piston. The initial 91 forced harmonic frequency sweeps were limited to the range from 20 to 22 kHz, the frequency space covered by the examined resonance. In order to avoid inaccuracies in the peak measurements due to the discretisation of the frequency axis, a set of finely resolved secondary frequency sweeps centred around the found resonance frequencies was added. The above histogram lists the design parameters (their meaning can be inferred from figure 5.2, p. 128) along the horizontal axis and the percentage change of the pressure amplitude peaks on the vertical axis. The effect of decreasing a geometry parameter in three steps is visualised in the left block of bars, the effect of increasing steps in the right block. The colours blue, green, and brown represent the first, second, and third step in each direction. The inset illustrates the situation for the parameter *gap* by marking the values taken for the bar chart in a finer resolution variation sequence. The histogram shows that a majority of the parameters has a great impact on the sound pressure amplitude peak found within the examined frequency band.

The question of how much luck needs to be involved in order to produce a well-performing resonator (as may have happened at ORNL) can also be seen under a slightly different angle when considering that science needs independent verifications or falsifications to proceed. Any description of resonator construction and handling must be transferable from one team to another. If a recipe leads to a 30 % success rate in one team but only to 5 % success rate when passed along to a different team, then the situation is clearly sub-optimal.

Consequently, the basic goal of the collaborative efforts of the RPI and KIT teams became the design of a new well-performing and robust acoustic resonator. Here, the second insight from the sensitivity study became important, which was that two designs can only be compared after thorough parameter-tuning. The sound pressure maps, where resonances shift, where pressure antinodes rise and fall when design parameters are varied, revealed that the parameter-tuning for achieving maximal sound pressure gains represents a challenging optimisation problem because of the number of parameters involved and the fact that there are many local optima. As multiple design ideas and setup variants (e. g. switching between glass, aluminium, and steel for selected components) were to be examined using the FEM model, an automatic design optimisation routine was sought and a solution to the problem found in evolutionary algorithms (EA) employed as global search routines.

A preliminary decision to explore an approach of EA-based design optimisation yielded a decisive proof of concept. The FE model of the highly sensitive West-Howlett resonator was first optimised manually consuming a budget of two to three thousand forced-harmonic FEM simulations yielding a setup achieving a sound pressure of 22 bar. In comparison, a simple ad hoc implementation of a hybrid EA was able to achieve a sound pressure amplitude of 27 bar based on three optimisation runs with an overall budget of around 9000 solver calls. The simulation budget increase pales as a contra argument when compared to the reduction in human effort and the workflow speed-up realised when EA optimisation runs requiring little attention replace weeks of scripting, data, analysis, discussion, and decision making. After the first EA trials provided such a clear proof of concept, a plan was made to advance the sonofusion topic by proposing and investigating improved resonator designs based on a framework of automated EA optimisation of FE models.

The following chapters are to explain how the task of proposing new resonator concepts has been undertaken, they describe the resonator design problem, the comprehensive global and local optimisation approach, and, at the end, new design suggestions. Not only these final design instances, rather the whole resonator analysis, design, and optimisation approach is intended as a contribution to advance the science of sonofusion.

1.7 Summarising the status of SF research

Sonoluminescence is a very interesting phenomenon to study because of the extreme degree of energy concentration and the highly transient nature. Sonofusion would be a new relatively low-tech way to achieve controlled thermonuclear fusion reactions, and SF neutrons would represent a new and quantitatively precise probe into SL plasmas (motivation in terms of fundamental research). Whether SF could become

as interesting as other fusion techniques to be examined as a suitable way for energy harvesting is at this point not at all answerable[37] (motivation in terms of energy-related applied research). But given the principal low-tech nature of SL experiments and the conceptual simplicity of SF detection, it is definitely worthwhile to pursue the next steps from SL towards SF as long as there is no general acceptance of principally unsurmountable hurdles. This line of thought is reflected in SF-related research projects which are undertaken at this moment, e. g. by the company Burst Laboratories[38] which counts several of the world's leading SL researchers among its associates.

Experimental SL setups exist in many variations, e. g. SBSL, MBSL, aquæous or non-aquæous liquids, pressure waves in resonance or pulsed, and several routes are being followed to drive SL plasma to fusion conditions. However, the setup of Taleyarkhan et al. is the only one of which several published claims of observed fusion already exist. A plausibility backing exists as well in the form of the theoretical work of Nigmatulin et al. describing in detail important unique features of spherical bubble cluster implosions in chilled degassed acetone. The main problem of this branch of sonofusion research is its history which went from a short initial hype directly into oblivion. This awareness and status decay within the research community may be partially rooted in science-irrelevant surrounding facts ranging from US congressional politics down to the pettiness of competing research groups. Certainly it has a lot to do with the very science-relevant status of the condition of general independent verifiability which can be seen as ambiguous upon close look, but is simply negative when counting primarily peer-reviewed published texts and applying a strict and conservative judgement. Therefore, the current situation of this branch of SF research can be described as being in an unfortunate stall, whereby sound and independent replications of the experiment would be highly desirable in order to clear the ambiguity of the verification status.

1.8 A shift in focus and proposing a way forward

This dissertation intends to advance the stalled discussion by shedding new light onto the technical issues of the employed SF resonator designs. From such seeming details as

- the history of resonators investigated by the RPI team and their inconsistent performance,
- Ross Tessien abandoning this resonator technology after serious trials (being quoted with "It's a nightmare to run it, and it breaks"),

[37] Although it may be too early for answers, yet it is still possible to add key thoughts. On the contra side it can be reminded that as long as the working liquid has to be kept cooler than ambient temperature thermodynamics prevents efficient energy harvesting. On the pro side it can be said that standing wave patterns can be scaled up to form regular patterns (multiplication of active sites) and that liquid metals as working liquids have the advantage of combining high density with low vapour pressure and offer a way towards higher temperatures.

[38] Burst Laboratories Inc., Grass Valley, CA (USA), URL: `http://www.burstlabs.com/`

- the fact that only a small fraction of many resonators perused by Taleyarkhan et al. actually led to positive SF data [253], and
- the fact that overall only few SF demonstrations by Taleyarkhan et al. were documented, a sequence which unfortunately stopped several years ago,

and from the preceding RPI-KIT project of resonator characterisation and modelling it may be extrapolated that the difficult reproducibility of well-performing resonators may be a crucial factor, and a largely overlooked one as well. If this was the case and the stalling situation were to a significant part attributable to a severe reproducibility problem of the employed resonator design, then the proposition of improved design concepts would be a desirable key contribution to advance the field of SF research.

Finally, it can be stated in brief: the working hypothesis is that the resonator design having been used by Taleyarkhan et al. is too sensitive and affected by a severe reproducibility issue. On this basis an attempt to provide improved resonator designs and a design methodology is being made. After backing the working hypothesis by a resonator sensitivity study based on FEM simulations (documented in appendix Q) this attempt consists in proposing a simulation-aided resonator design methodology involving performant evolutionary global search algorithms and presenting exemplary optimised resonator designs.

1.9 Short description of chapters and appendices

1.9.1 Main body chapters

This thesis and project documentation consists of a relatively tight main part and an additional compound of appendix chapters. The topic structure of the four main chapters is the following:

- Chapter 2 outlines the task of designing liquid-filled acoustic resonators for SF experiments from the bottom up, from the fundamental physical requirements for the sound field to the demands arising from a real-world laboratory situation. It concludes with rephrasing a central part of the engineering task as a function minimisation problem.
- After briefly outlining the principles of evolutionary algorithms chapter 3 presents the development and benchmarking of the newly developed hybrid EA.
- In order to turn an engineering problem into a function minimisation problem, a design evaluation routine is needed which postprocesses simulation outputs and serves as a fitness function associating each solution candidate with a scalar quality measure. Chapter 4 describes the logic behind the different fitness evaluation routines employed during EA optimisations of SF resonators.
- Finally, chapter 5 presents the geometries proposed here for improving the basis of SF experiments. It consists of a compilation of EA optimisation case studies. The case studies highlight technical aspects, methodological questions, and the properties of the optimised resonator setups. Some of the case

studies illustrate general challenges arising when applying EAs to real-world optimisation problems.

Due to their methodological focus chapters 3, 4, and conditionally 5 may find the interest of a more general readership whereas the other chapters of the main section belong to the topic of SF. Readers purely interested in the discussion of SF and resonators can leave out chapters 3 and 4.

1.9.2 Appendix chapters

The compound of appendix chapters has three purposes. Firstly, one aim is to supply some context allowing an easy and quick access to the topic and literature of sonofusion for interested readers. SF experiments represent a region of overlap of the topics of nuclear physics, nuclear instrumentation, the thermodynamics of liquids, vapours, gasses, and plasmas, and finally acoustics, structural mechanics, and electrical engineering. The inconclusive state of affairs after several replication trials by scattered teams around experts of sonoluminescence or nuclear engineering may perhaps be interpreted as a consequence of the difficulty of illuminating all the multi-physical aspects of the phenomenon with expert understanding at the same time.

The second purpose is to supply beneficial background knowledge for prospective applicants of EA-based optimisation workflows to various optimisation problems in research, engineering, or industry. Educated users can make better decisions on choosing and modifying algorithms, and effective users of global optimisation workflows can speed up their cycle of numerical experiment, analysis, and understanding. The minimisation of computational budgets through intelligent algorithm choices allows to achieve a further speed-up.

The third purpose is to document the technical aspects and main results of the RPI-KIT collaborative work on sonofusion which included an experimental campaign of resonator characterisation and the development and investigation of FEM simulations.

In terms of topic relations an appendix chapter group listing can be made in the following way:

- Overlapping disciplines – introduction: **A**
- Overlapping disciplines – physics of fluids: **C, H**
- Overlapping disciplines – plasma & nuclear physics: **B, E, F, G**
- Overlapping disciplines – electromechanical transducers: **J, K, N**
- FEM simulations of liquid-filled resonators: **K, Q, R**
- RPI & RPI-KIT SF-related activity documentation: **D, I, O, P, Q, R, S**
- Evolutionary algorithms: **T, U, V**

Furthermore, the appendix chapters can be divided into a first block (A-N) serving the function of a comprehensive literature and situation review and a second block (O-W) serving a documentary function with respect to the cooperative RPI-KIT contributions on the topic of SF.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
a	acceleration
c	speed of sound/light
E	energy
F	force
I	current
M	Mach number
m	mass
N	count number
p	pressure
Q	quality ("pointedness" of a resonance peak)
R	radius
T	temperature
U	voltage
v	velocity

List of Greek quantity symbols

Symbol	Description
σ	surface tension; standard deviation

List of particle symbols

Symbol	Particle
D	deuteron $= p + n$
e^-, e^+	electron, positron
p	proton
n	neutron
$\nu_e, \bar{\nu}_e$	electron neutrino, electron antineutrino
T	triton $= p + 2n$

List of abbreviations

Abbreviation	Description
AFTEC	Acoustic Fusion Technology Energy Consortium
AICF	acoustic inertial confinement fusion
APDL	Ansys Parametric Design Language
BBC	British Broadcasting Corporation
CFR	cavitation fusion reactor
CV	coefficient of variance
DARPA	Defense Advanced Research Projects Agency
EA	evolutionary algorithm
FE,FEM	finite element (method)
HBT	Hanbury-Brown Twiss
IKET	Institute for Nuclear and Energy Technologies (Institut für Kern- und Energietechnik)
KIT	Karlsruhe Institute of Technology (Karlsruher Institut für Technologie)
LS	liquid scintillator
MBSL	multi-bubble sonoluminescence
MD	molecular dynamics
NED	Nuclear Engineering and Design (journal)
NURETH	International Topical Meeting on Nuclear Reactor Thermal Hydraulics (conference)
ORNL	Oak Ridge National Laboratory
PM,PMT	photon multiplier (tube)
PNG	pulsed neutron generator
PSD	pulse shape discrimination
QED	quantum electrodynamics
RP	Rayleigh-Plesset
RPI	Rensselaer Polytechnic Institute
SBSL	single-bubble sonoluminescence
SF	sonofusion
SL	sonoluminescence
SLAC	Stanford Linear Accelerator Center
TCSPC	time-correlated single photon counting
UCLA	University of California, Los Angeles

Chapter 2

The sonofusion resonator design problem

Before thinking about modified or alternative SF resonator designs, this chapter lays out a systematic analysis of the SF resonator design problem. On the one hand the requirements for the functioning of a cavitation-enabling acoustic resonator are discussed on a fundamental level. On the other hand the list of requirements is made comprehensive by incorporating lessons learnt from past SF experiments. The main function of the chapter is the ordering of thoughts leading to new SF resonator design proposals. Another purpose is to lay the foundation which will later allow the discussion and judgement of detrimental and beneficial objective function variants.

The reproducibility issue of the acoustic resonators of the type used by Taleyarkhan et al. arises from the conflict between the sensitivity of the design and the large tolerances being associated with the manual work steps in the manufacturing process such as glassblowing or assembly without precision gauges (holding frames, jigs). The resolution of the reproducibility issue can be found either by identifying a more robust (less sensitive) design or by transitioning to other manufacturing techniques with well controlled tolerances. In principal, robustness evaluations require lots of design variation evaluations as soon as one intends to go beyond local gradient measurements. Optimisation strategies with repeated intermittent sensitivity or robustness evaluations will always involve substantially elevated computational budgets as compared to pursuing straightforward optimisation approaches. The even larger problem is the principal uncertainty at the outset whether substantially more robust design points with equal or better performance than the reference case can be found at all. Thus, design ideas allowing different manufacturing techniques were pursued.

Glassblowing is an ideal technique for producing hollow shapes. In the acoustic resonator design of C. West and R. Howlett adapted for SF trials by Taleyarkhan et al. hollow structures are used for the main section as well as for the pistons. Due to the elasticity of glass these forms offer vibration mode shapes with vast areas of large normal displacement amplitudes which is crucial when serving as boundary

conditions for the fluid domain of an acoustic resonator. Classic precision machining techniques are much less suitable for creating closed or almost closed hollow shapes. When intending to suggest equally performing resonator designs based on precision-machined parts (of glass or metal) the difficulty lies in negotiating different goals, namely:

- that resonator parts must be assembled to form a closed volume,
- that the resonator structure be flexible but not prone to breakage or dissipation-induced heating,
- that neither glued assembly bonds nor tubing connections introduce excessive levels of damping or an excessive risk of breakage, and
- that acoustic modes with a desirable sound field topology can be established.

In a practical resonator design structural shapes, manufacturing techniques, and easy assembly are required, but for a well-performing resonator the requirements are much higher. In a high-quality resonator the motion patterns, masses, and stiffnesses have to fit together perfectly which involves not only the outer structure, but also the inner fluid volume and the coupling properties. This is why the resonator design problem consists of two parts, the actual design problem and the subsequent proportion tuning task. The following chapter serves to lay out both these tasks in a fundamental way.

2.1 Requirements of a functional and reproducible resonator

With the intention of writing down a bottom-up list of requirements for a well-performing resonator design, let us begin with the basic relationship between the shape and possible standing sound waves in a fluid volume. The equation describing acoustic waves is the same in a gas or liquid, but let's focus our imagination on a liquid because it can bear states of tension. Unlike a gas it can store energy through both compression and tension which is in fact the case in resonators used in the past for SL or SF experiments. A spherical drop of liquid hovering freely in zero gravity[1] and oscillating in one of its eigenmodes is in theory the best SF resonator because it is wrapped in a zero pressure amplitude boundary condition (BC) all around. In the fundamental eigenmode all motion is in the radial direction and with an amplitude decreasing towards the centre. The zero pressure amplitude and maximum displacement BC on the bubble surface is mirrored by a zero displacement and maximum pressure amplitude condition in the centre. Closest to that in an earth-bound lab setup comes a thin-walled spherical glass flask filled with liquid and excited to oscillate in the fundamental radial mode, as symbolically sketched in figure 2.1.

[1] Shock waves and cavity implosions in a spherical drop of water floating in zero gravity are actually possible, see [331] and `http://www.youtube.com/watch?v=zpkVN64GIt4`.

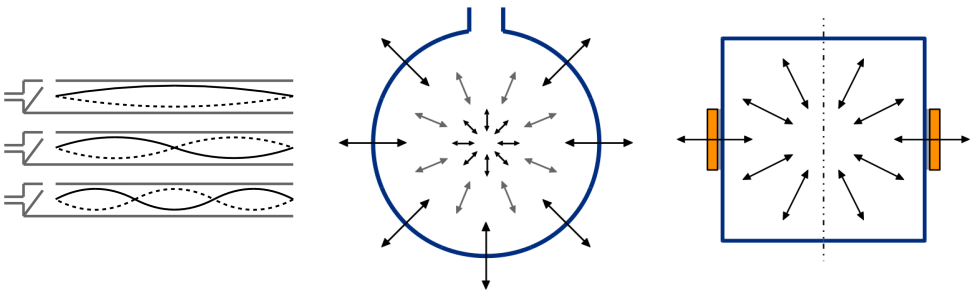


Figure 2.1 Ideal geometries of the liquid domain
The simplest case of a standing pressure wave encountered in everyday life is probably the organ pipe or flute. The longitudinal standing wave inside represents a 1D problem. The first three pressure eigenmodes of the labial pipe (flue pipe) are depicted on the left. The pipe has two open ends determining the location of pressure nodes (motion antinodes) [38]. Inside the pipe N pressure antinodes (motion nodes) can be located. If a spherical droplet of liquid or a ball of elastic material is excited to oscillate in its fundamental radial eigenmode, then there is a pressure node all across its free surface and a motion node in its very centre. This means the largest radial displacement amplitude is found at the surface. Containing a spherical volume of liquid in a glass flask, its radial modes can be excited if the flask wall can be put into contraction and expansion motion (see centre sketch). The length of the arrows symbolises the motion amplitude decreasing towards the centre. Ideally, the glass wall should be as thin as possible in order to influence the eigenmode of the fluid as little as possible, the glass surface should be covered homogeneously with tiny piezoelectric transducer rings letting the glass layer contract like a skin; and the nozzle should be small. But already very simple setups with two transducer rings glued on a common laboratory flask allow to create sonoluminescence in water [501] even though the symmetry is not perfect and flat transducer rings have to be glued onto the curved surface. The next step towards a resonator with easier controllable part machining and assembly techniques is to use a segment of an industrially manufactured glass tube the ends of which are sealed with flat plates. The coupling to the transducer becomes easier as flat piezo rings can then be glued onto flat end plates [370] or hollow cylinders around the glass tube, in both cases with a homogeneous glue layer thickness. Moreover, the transducer does not further reduce the symmetry. However, large bending at the edges is unavoidable and must be dealt with because side walls and end plates have to move in- and outwards in phase.

As soon as there is a hull of solid material around the liquid, the pressure amplitude generally deviates from zero at the interface. Some forces have to be transmitted so the motion of the solid can excite the motion of the liquid. Secondly, if the resonance frequency of the hull is not exactly the same as for the liquid's eigenmode, forces have to keep the motions in lockstep. This explains why the inner radius of the glass sphere does not necessarily have to coincide with the radius of a pressure node and a motion antinode of the targeted eigenmode of the desired fluid volume. It also explains what the benefits of numerical simulations can be, even in this case of a trivial geometry, because the approach of analytically matching resonator dimensions with sound wavelengths in the liquid is too simplistic. It's the eigenmodes of the coupled system of structure and liquid that count.

As noted in figure 2.1, going from spherical to axial symmetry, the next geometry is that of a cylindrical fluid volume. Also a cylinder-shaped chunk of liquid or elastic solid can vibrate in its fundamental radial eigenmode. The problem arising for the structural hull are bending stresses and fatigue at the cylinder edges where the angle between side wall and end plate has to widen and narrow when all walls move in- and outwards in phase. Nevertheless, the cylindrical shape is the geometry used in most lab setups for creating sonoluminescence. Figure 2.2 shows some options of how to deal with the bending stress problem.

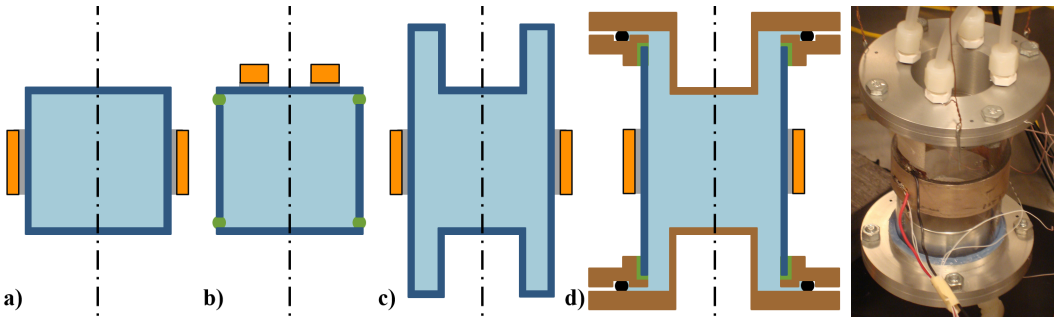


Figure 2.2 The problem of bending corners in cylindrical resonators
In order to overcome the situation with stress peaks in the corners (a), connections via a sufficiently flexible glue can be used (b). Sketches (a) & (b) also show the two alternatives of transducer positioning: as a hollow cylinder glued around the glass cylinder or as a ring on top of it. With the new resonator design (c) proposed by Cancelos [69] the RPI team was aiming at giving the glass tube segment more way on both ends for the displacement amplitude to decay, thus reducing bending angles and stress levels (see also resonator 7 in fig. I.5). The structure to the right (d) encloses exactly the same fluid volume as its neighbour and has the same cylinder wall extensions. It is the variant with bolted aluminium flanges designed by Cancelos. The metal flanges in conjunction with silicone glue connections between metal and glass are to solve the problems of excessive stresses and fatigue.

Within the sequence of SF-related projects the RPI team tried options (c) & (d) (resonators 7 & 8 of in fig. I.5) for building well-performing resonators achieving high cavitation rates in acetone, but both approaches failed: the all-glass design cracked due to fatigue, and the aluminium flange design exhibited a low Q-factor in connection with significant dissipation-induced heating such that the acetone began to boil during cavitation. The analysis of the design had then been backed up by FEM simulations dring the RPI-KIT cooperation project (see appendix Q). Trying to learn from these experiences, the advantages and disadvantages should be listed:

- **Resonators of the West-Howlett style:** Their advantages are a high Q-factor and a high sound pressure amplitude. The FEM simulations show that in the right setup the latter is connected to a displacement amplification mechanism along the main glass wall. The great disadvantage is the sensitivity with respect to component dimensions and the ensuing conflict with the large tolerances due to manual glassblowing as the principal manufacturing technique. Indeed, one must be very lucky to obtain a well-performing resonator. A further disadvantage of the existing resonators of that design is the low number of outlets and that they are only placed on the top. If the top head is connected by a silicone bead and if this bead has to be cut open for refilling, then the setup is not reproducible after refilling. If the connection is by flange, then the resonator cannot be reliably simulated, and the mechanical coupling across the flange might not be reproducible either.
- **Resonators of the Cancelos style (H-formed cross section), all glass:** One design tried at RPI (resonator 7) exhibited a high Q-factor and sound pressure gain [392] but cracked due to fatigue in the region of high glass wall curvature at the top rim. Another disadvantage is that the manufacturing involves relatively complicated glassblowing work.
- **Resonators of the Cancelos style (H-formed cross section), with**

flanges and metal pistons: The great advantages of that design are that all the metal parts can be made with CNC machining, that the glass part can be cut from an industrially manufactured glass tube, that gauges (setting jigs) can be used for the assembly by glueing, that small holes can easily be drilled into the metal end plates and equipped with teflon nozzles, so that enough nozzles can be placed for filling and draining without opening the resonator, that the nozzles become an ever smaller perturbation the more massive the end plates are, and that the flanges make assembly more reproducible. The great disadvantages are the bad cavitation performance because of too high damping, dissipation-induced heating, and acetone boiling, and that the bolted flange connections make it hard to simulate this resonator.

After the learning process consisting on the one hand of the experiments by the team at RPI who built and used various resonator versions and on the other hand of the thorough resonator characterisation measurements and FEM simulations of the RPI-KIT collaboration project, an updated list of general requirements can be formulated for any new resonator design. These are the primary challenges of the sonofusion resonator design problem:

- The parts and the whole resonator assembly must be reproducible, a resonator should deviate in its sound pressure performance only little from what is expected from a given design based on experience and simulation.

- The resonator needs to create a good sound field. This means the antinode with the highest sound pressure amplitude has to be located in the central region. This is where spherical bubble clouds are to appear and implode. Pressure amplitudes near structural surfaces should be substantially smaller because cavitation on walls is highly undesired. As the bubble clouds are desired to collapse due to a spherical implosion front propagating towards the centre, the central sound pressure antinode should coincide with a motion node. The pressure and displacement fields in the whole central area should show spherical symmetry, i. e. the symmetry break due to the outer structure of the resonator should become negligible in the central region far away from the walls.

- The structure should vibrate in a manner so there are no points of stress concentration. Breakage at chamfers and due to fatigue must be avoided.

- The structure should efficiently couple the vibration of the transducer with the desired oscillation mode of the fluid.

- The resonator should exhibit low damping and a high Q-factor.

- The resonator needs cooling. If the damping ratio is elevated, the resonator design has to comprise a sufficient cooling concept.

- A sufficient number of outlets and nozzles is required for filling, draining, and controlling the liquid level during experiment campaigns. Some nozzles should

be placed where they can serve as outlets for eventually created noncondensible bubbles. The nozzles and connected tubing should not deteriorate the resonator's vibration and damping behaviour, they should not act as a bridge for letting kinetic energy leak out of the vibrating system.

- As experience with the West-Howlett design and its associated FEM simulation shows, a mechanical amplification of the transducer motion by the structure can become an important design feature.[2]

2.2 Comparing resonator designs

A comparison of resonator designs or resonator design concepts has to be primarily based on the fulfilment of the requirements listed above, but as soon as two designs fulfil those conditions, the more interesting question of the resonator performance comes into play. A resonator performs well if the sound pressure amplitude is large, if the region of largest amplitude is large and spherical, and if the sound pressure amplitude elsewhere is low, particularly near the walls. A resonator performs well if it doesn't require an excessively complicated or large transducer to power it. A possible straightforward definition of a quality measure is the sound pressure amplitude per transducer driving voltage, also called the *pressure gain*. Instead of the input voltage a similar measure could be based on the input power.

Characteristic measures like the sound pressure amplitude or gain at a given location are frequency-dependent numbers which are by definition diverging or peaking at resonance frequencies. Clearly, a comparison of two resonators by pressure gain makes only sense if both are tuned to run at resonance. Oscillating systems are determined by masses and stiffnesses, and variations of masses and stiffnesses can detune a resonator similarly to a change in excitation frequency. This is the simple reason why an objective comparison of resonator designs can only be made after ensuring that all dimensions, masses, stiffnesses are adjusted so that a quality measure like the pressure gain is maximised for a given resonance mode. This defines a parameter tuning task, i. e. an optimisation problem. In a practical approach the range of possible and potentially good solutions can be widened by allowing the resonance peak to be positioned within a wide frequency band instead of insisting on a predefined target frequency.

If an objective comparison between resonator designs requires well-tuned, i. e. optimised, instances of the designs, then the next question is about how to tune a given design. Looking at the example of organ pipes shows that in this case tuning means mainly sizing. In a simple SF resonator geometry like a spherical or cylindrical

[2]The FEM simulations show at least three distinct resonances for the frequency range in question. One of them exhibits a strong pressure antinode in the liquid located on the central axis well above the PZT transducer ring. This location coincides with the descriptions given by Taleyarkhan's group and Saglime (RPI) of the sites where bubble clusters occur. The simulation reveals a particular feature of this mode shape: the cylindrical glass wall has several radial displacement nodes and antinodes along the vertical axis, and the next antinode above the transducer has a much larger amplitude than the antinode directly behind the transducer. Thus one can speak of a mechanical amplification of the displacement by the vibration mode shape of the glass wall. This can be seen in plot Q.9 (p. 439) in appendix Q.

thin-walled glass flask one could try to proceed analoguely by calculating the right dimensions for a given frequency based on how the liquid's bulk modulus determines the connection between wavelengths and frequencies. But as already discussed, the assumption of an extremely thin-walled flask with negligible influence of the structure on the oscillation modes might be too simplistic. It is fluid-structure interaction which has to feed the acoustic field in the liquid with energy from the transducer. Taking account of the structure will complicate the situation, in particular for arbitrary geometries. This is why simplistic resonator layout calculations will not do the job. A simulation-based approach of resonator tuning through systematic design variation can be much more efficient. The immediate next arising question is then about the characterisation of the resonator optimisation problem and whether it can or should be addressed by manual or automated algorithmic approaches.

2.3 The finite element model at the basis of resonator comparisons and optimisations

Finite element (FE[3]) models of the acoustic resonators have been used in this study as the main tool to achieve a deepened system understanding by accompanying the experimental resonator characterisation campaign and by conducting parameter variation studies. During the RPI-KIT collaboration project several 2D-axis-symmetric FE models have been composed in the FEM software suite ANSYS® using the scripting interface APDL (ANSYS Parametric Design Language) and benefiting from dedicated sections of the element library for including the properties of piezoelectric materials which couple the mechanical with the electromagnetic domain. The modelling approach has been benchmarked against older FEM simulations which had been conducted by Cancelos at RPI [69] and the database resulting from the latest experimental campaign of transducer and resonator characterisation which was part of the RPI-KIT collaboration. This work[4] was published at several conferences together with B. A. Malouin, R. T. Lahey Jr., A. G. Class, and T. Schulenberg [433, 435–437]. The models and the investigation results are documented in much more depth in appendix chapter Q. The key insights which were already mentioned in a loose anecdotal form in chapter 1.6 can be summarised in the following short fact list:

- The FE model of the West-Howlett style SF resonator is able to produce a sound pressure map (amplitude and phase data over frequency and axial position) which qualitatively matches experimental benchmarking data gathered by hydrophone.

- The simulations yielded a clear picture of the mode shapes exhibited by this resonator design and allowed a better understanding than was documented in previous publications on SF resonators.

[3] alternative abbreviation: FEM for *finite element method*

[4] All experimental work was a collaborative effort shared among B. A. M. and M. J. S. whereas the FEM simulations were set up and conducted by this author.

- The resonator performance is very sensitive to a majority of the design parameters.
- The extreme pressure gain sensitivity with respect to some key parameters where fractions of millimetres can make a difference suggests that due to manual glassblowing being involved in the manufacturing process this resonator design is not reproducible for robust and reliable performance.
- The high degree of sensitivity also suggests that deviations from the axial symmetry of the real-world resonator can prevent a perfect match of the pressure maps from simulation and measurement.
- The design parameters with high sensitivity are coupled because various spatial dimensions, masses, and stiffnesses have to be matched in order to yield optimal pressure gain performance.
- In certain cases the resonator wall at some distance away from the transducer can exhibit a much larger displacement amplitude than found directly behind the transducer. The mechanical displacement amplification mechanism may be important for yielding optimal sound pressure fields.

Thus, the shortcomings of the old resonator design and the need for new design concepts was revealed, while at the same time enough information could be gathered which would allow subjecting the scripted and parametrised FE models to algorithmic blackbox optimisation procedures. The fact that for non-trivial resonator geometries there exist several resonances within the promising frequency band and that the resonances shift, grow, and decay as parameters are modified rules out unsystematic manual tuning approaches.

2.4 Resonator-tuning as an optimisation problem

Resonator tuning means finding a set of design parameters which maximises a response function like the acoustic pressure, pressure gain, or a similar derived function. Such minimisation or maximisation tasks related to a function $f(\vec{x}) : \mathbb{R}^n \to \mathbb{R}$ are in principle search tasks: the search for the highest summit or the deepest valley is the search for the right spot in the search domain $\mathbb{R}^n$, and whether this search task is challenging or easy is mainly determined by the dimensionality and topology of the objective function $f(\vec{x})$.

How hard can the optimisation problem of SF resonator tuning be? The sizing of a spherical glass flask for a given working liquid while neglecting any outlets or the transducer would involve only two parameters, the radius and the wall thickness. However, the consideration of design parameters for the outlets and transducers already leads to many new design parameters. The transition away from the trivial geometries of sphere and cylinder towards more complex ones also has the consequence of increased sets of design parameters. The simulations and the sensitivity study discussed in appendix Q show that among dozens of design parameters a majority can be expected to play a substantial role.

After discussing the dimensionality, what can be said about the topology? Keeping the driving voltage amplitude fixed in the FEM simulations, the pressure gain and the pressure amplitude are two equivalent measures which can be taken as the primary objective function. The fact that there can be several resonance peaks with different mode shapes in the investigated frequency interval and that the pressure peaks rise and fall when parameters are varied means that there are local optima separated by design space areas representing mediocre design points. As a direct consequence of this "competition of many resonances" there are many local maxima of the objective function separated by valleys of worse acoustic pressure performance. The other way to think of it is that there are many possible combinations of masses and stiffnesses allowing the assembly of structural parts and the fluid to oscillate in a mode with a strong pressure amplitude. If a stiffness is changed here, then a mass needs to be re-tuned there, in order to make an incremental design step without deteriorating the system performance. Due to these correlations, the problem is nonseparable. And because the resonances interact in various manners, not only with overlapping and multiplication when one resonance crosses another but also with repulsion and quenching, because of these properties it is also clear that the search for the largest acoustic pressure amplitude is not as simple as following the gradient and a branched structure of ridgelines (shaped like a tree diagram) to the one global peak.

The investigations of the resonator FE models during the RPI-KIT collaboration project, in particular the sensitivity study of the pre-optimised West-Howlett style resonator in the closed SF setup are able to verify this description of the optimisation problem. The main results are documented in appendix Q. The figures which are particularly suitable for illustrating the "competition of many resonances" are Q.20 and Q.19.

The sound field quality is not only determined by the pressure gain. A secondary goal is a low maximal acoustic pressure amplitude along walls. Cavitation near the walls is undesired as it destroys the desired acoustic field intended to enable the spherical implosion of bubble clusters. This secondary goal defines a constrained optimisation problem: maximise the pressure gain while keeping the pressure amplitude near walls below a certain threshold. It should be expected that such a constraint will lead to the rejection of a subset of resonances in a subset of the investigated search space. A statement whether this makes the optimisation task more or less challenging can a priori not be made. One way to prevent undesired designs as the outcome of an optimisation process is to manually impose sufficiently narrow bounds for the tuning parameters in an optimisation process under tight human control. This will generally make the scanned portion of the design space smaller which contradicts the wish to explore diverse design variants as much as possible.

Summarising the task and proposing a solution approach

The characterisation of the optimisation task of SF resonator tuning can be summarised a bit more concisely: when the design parameters are varied, resonances rise and fall, they shift in frequency, they cross, overlap, and quench each other, they enter and leave the scanned frequency interval. Therefore, the resulting search

landscape is multi-dimensional (5-30 parameters to tune), smooth, but it has many irregularly distributed local optima (i. e. acoustic pressure maxima) separated by valleys of various depths. By consequence, the task can be classified as a hard[5] global optimisation problem such as e. g. optimising a spacecraft trajectory including several swing-by manoeuvres (multiple gravity assist, MGA). A common approach to address this class of optimisation problems is the utilisation of *evolutionary algorithms (EA)*. EAs are stochastic search algorithms which are inspired by the concept of gene pool improvement through the principle of "survival of the fittest" in combination with search space exploration based on stochastic operators such as mutation and recombination operators. Just as biological evolution acts on construction plans of living beings coded in the four-letter alphabet of DNA and RNA, EAs act on populations of solution candidates represented by lists of numbers (i. e., vectors) or symbols (e. g. the binary alphabet of computers or symbols with special meaning in a construction plan or syntax for generating candidate solutions).

The feasibility of tuning resonator FEM models with EAs

Due to the concepts of randomised global search and information accumulation within a gene pool EAs run more efficiently the more design variants can be evaluated. Computational budgets ranging from several dozens to many thousands of design variant evaluations are common for EA applications. The investigated SF res-

[5]Instead of the soft classification "hard" the field of complexity theory offers the sharper classification "NP-hard", whereby "NP" stands for "nondeterministic polynomial (time)" [157]. The term is often mentioned in the context of EA application cases as it is commonly accepted that NP-hardness implies *practical intractability* and justifies favouring approximative methods or heuristic approaches over exact and exhaustive algorithms for finding provably optimal solutions. The description as NP-hard means that the problem in a generalised form requires an algorithm for its solution for which the computation (time) budget grows at least polynomially with problem size under the hypothetical assumption of using a nondeterministic computer ("oracle machine"). This may imply more than polynomial (i. e. exponential) growth with real-world computers. The necessary definitions and proofs of NP-hardness are mostly restricted to decision problems (answer is "yes" or "no"; e. g. can a given partially filled Latin square be completed (Sudoku) [90]? is a given Rubik's cube pattern solvable in less than k moves [228]?) and they are commonly based on carefully chosen symbolic problem instance encryptions and a translation of the problem class into graph theoretical terms [157]. Problems with strongly combinatorial character like scheduling problems, logic games, and puzzles offer themselves much more straightforwardly in terms of complexity and intractability analysis as many function optimisation and engineering problems. The problem of multi-gravity assist trajectory planning was described as NP-hard [486] owing to the fact that the number of possible paths grows exponentially with the number of celestial bodies [77]. The MGA problem exhibits a similarity with a magnetic pendulum, a famous example of deterministic chaos, that – not everywhere in the parameter space, but in many places – tiny trajectory deviations can lead to completely different final destinations. Bifurcations separate patches in the parameter space, all points within a patch leading to the same destination. This is because in a swing-by manoeuvre small variations in the periapsis cause large deviations in the angle of the outgoing branch. By consequence, combined recipes of branch exploration and local optimisation yield efficient MGA planning algorithms [487]. Suffice it to say that a two-fold similarity can be observed between MGA planning and SF resonator design: (a) as the number of celestial bodies drives the amount of MGA path options, the number of masses, springs, and geometry details drives the number of possible motion patterns of a vibrating structure, and (b) because of the "competition of resonances" the resonator tuning task exhibits a patch-wise smooth objective function in the form of $p_{\max}$ where patch-internal local search means pushing up the pressure peak of a given antinode belonging to a given mode shape.

onator FE models are 2D-axis-symmetric FE meshes comprising several thousand nodes. The frequency response of a resonator is gained through a forced-harmonic analysis which should in principle cover the relevant frequency interval (several kHz) with a very fine resolution due to the sharp resonance peaks. Computation time can be saved by scanning in steps, first coarsely, then with high resolution only in the most interesting frequency bands. This way a single design evaluation requires only a few minutes on a contemporary personal computer. With some degree of parallelisation and conducted on an institutional computer cluster the optimisation runs will take not more than a few days. This means the global optimisation of many FE model variants is feasible within a time frame of a few months. The next step is the choice of a suitable optimisation algorithm. The more efficiently the global search task is handled, the more different geometries and setups can be investigated within the available time.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
N	count number
Q	quality (“pointedness” of a resonance peak)
$\mathbb{R}$	real numbers
$\vec{x}$	point in search space, set of design parameters

List of abbreviations

Symbol	Particle
APDL	Ansys Parametric Design Language
BC	boundary condition
CNC	computerised numerical control
DNA	deoxyribonucleic acid
EA	evolutionary algorithm
FE,FEM	finite element (method)
MGA	multiple gravity assist, a deep space trajectory with swing-by manoeuvres
NP	nondeterministic polynomial
KIT	Karlsruhe Institute of Technology (Karlsruher Institut für Technologie)
RNA	ribonucleic acid
RPI	Rensselaer Polytechnic Institute
PZT	lead zirconate titanate (a piezoelectric ceramic)
SF	sonofusion
SL	sonoluminescence

Chapter 3

Determining a hybrid EA scheme for resonator optimisation

Having made the decision that the challenging resonator optimisation task is to be addressed by applying evolutionary algorithms (EA) which rank among state-of-the-art global function minimiser algorithms, the question is which type of EA is the most suitable. It has been tried to make an objective decision by composing a telling selection of test problems and benchmarking some state-of-the-art EAs with it. An own hybrid EA scheme was added into the contest because the combination of several fundamental EA concepts promises performance robustness. This chapter describes the choice of algorithms, the selection of test problems, the development of the new hybrid EA, the benchmarking methodology, and the test results.

Optimisation algorithms and in particular EAs have in the past years become a vivid field of research and development. Many algorithms are documented in the form of articles, published source code, or open-source collaborative programming projects. Others are part of proprietary optimisation software packages. The selection of one or a few algorithms to apply to the problem of interest has to be made carefully. Even if a large number of simulations can be done easily in the case of the SF resonator FE model, the total number of carried out optimisation runs will be very limited as each run will consume hundreds or thousands of evaluations. The few affordable optimisation runs should be conducted only with the most promising optimisers.

It is common to evaluate the performance of EAs and other optimisers on dedicated libraries of test functions. Test problems are generally not computationally demanding as to allow many evaluations and the build-up of score histograms within short time. The EA testbed developed within this project does not only comprise test problems but also a code library of subroutines for coding EAs in the programming language Python. The initial intention was to be able to evaluate popular EA concepts neutrally, i. e. not dependent from implementation details. The byproduct

was the development of a performant hybrid EA.

The following chapter introduces some terminology of the field of EA, describes the developed hybrid EA, and presents the results of benchmarking it in comparison with other state-of-the-art EAs on a deliberately chosen collection of test problems. An additional background and context chapter on the field of EA including more terminology, basic concept explanations, and also covering some aspects of how advancements in evolutionary computation (EC) were and are able to deepen our understanding of the theory of evolution in general can be found by the interested reader in appendix T.

3.1 EA vocabulary in brief

The basic problem statement which can be addressed with evolutionary algorithms is:

$$\begin{aligned} \text{minimise} \qquad & f_{\text{obj}}(\vec{x}) : \mathbb{R}^n \to \mathbb{R} \\ \text{with} \qquad & x_i \in [a_i, b_i], \quad i = 1, ..., n, \end{aligned} \tag{3.1}$$

where f_{obj} is the objective function (or fitness function) and the intervals $[a_i, b_i]$ define the bounds of the search space. Solution vectors are in that context often called chromosomes. EAs are cyclic algorithms which mutate and update chromosome populations based on the principle of "survival of the fittest". A description within an operator-based conceptual framework can be very efficient and in these words an EA is a routine generating ever new offspring populations based on the application of operators with the tasks of *selection*, *mutation*, and *recombination* to a repeatedly updated parent population. A generic EA can be represented as a small snippet of pseudo-code:

Algorithm 1: generic generation loop of an EA

```
P ⟵ GenerateRandomPopulation(N)
f_obj ⟵ Zeros(N)
while g <= G and not StopCriterion() do
    for j ⟵ 1 to N do
        f_obj[j] ⟵ Evaluate(P[j])
    P ⟵ SelectionMutationRecombination(P, f_obj)
```

Transferring concepts from biology to applied mathematics: When talking about optimisation in the context of engineering problems, then the parameters x_i are system inputs or settings whereas a quantity $f(\vec{x})$ is an output ore response function. In the EA context the vector components x_i can be called *genes* because they are the meaningful elementary information-carrying entities which are mutated. Many more terms can be borrowed straightforwardly from biology to make EA descriptions concise and easily understandable. A solution candidate $\vec{x}$ can be called *chromosome*. Whereas a gene can be only subject to a mutation operator, a chromosome can be subject to both random mutation or recombination with another

chromosome. Popular examples of recombination operators are *single-point-*, *two-point-*, and *multi-point-crossover*; here, not only the word[1] but the whole concept is borrowed from biology. On the level of populations, terms like *gene pool diversity*, *selection pressure*, or *population bottleneck* can be usefully carried over into the EA context with barely any change of meaning.

μ-λ notation: There exists a practical shorthand notation for indicating the population renewal cycle. In a (μ, λ)-EA λ offspring chromosomes are created from a parent population comprising μ chromosomes. If $\mu = \lambda = 8$ then eight offspring vectors are created based on a parent gene pool of eight chromosomes. To close the generational loop, the offspring generation is called the new parent population, it replaces them completely. An example for the case $\mu \neq \lambda$ would be a (4,12)-EA: here only the four best chromosomes are selected from 12 offspring to form the new parent population. In the μ-λ notation a "+" sign can be used instead of the comma. The meaning of the term (4+12)-EA is that 12 offspring are created from 4 parents, and that the next parent population is composed by selecting the four chromosomes with the best fitness from the joined set of 16 chromosomes.

Constraint handling: The EA framework allows to circumvent complications arising from constraint functions often associated with optimisation problems and to keep the conceptual side simple by indirectly enforcing constraints via penalties on the fitness function.

A short description of popular EAs

The most popular EAs are genetic algorithms (GA), evolution strategies (ES), differential evolution (DE), and particle swarm optimisation (PSO). Swarm algorithms, in particlar PSO, are often classified as different from EAs but joined under the unifying roof label *nature-inspired algorithms*, although a broad enough definition of chromosome operators (and the term evolution[2] itself) allows the inclusion of PSO within EAs. In a classic GA, always two parent chromosomes are taken as input for generating one or two offspring chromosomes by swapping snippets (recombination) and randomly modifying single genes with a low probability (mutation). Originally, GAs were invented in connection with binary encoding of chromosomes; real-coded GAs were developed later. The development of ES was largely driven by thinking about optimal spatial distributions and distribution updates of point clouds for performing an efficient search. A single selected best chromosome or the mean vector of a few chromosomes forms the anchor point, i. e. the centre of mass, for the next offspring population which will be created by sampling from a univariate or multivariate normal distribution. DE is also based on thinking about mutation moves in the search space: the point cloud representing the current population is made to stretch out like an amoeba by adding scaled difference vectors found within the cloud to its tip. On top of that, GA-like recombination operators play a role. In PSO the space is searched along particle trajectories dominated by inertia and inter-particle

[1] The EA term "crossover" is short for "chromosomal crossover" or "crossing over"; the two latter terms are commonly used in biology where they describe the cutting and exchange of snippets between homologous chromosomes during meiosis.

[2] e. g. reproduction, mutation, and remanifestation of information in competition: in that framework pieces of information compete in PSO for being communicated to different particles

attraction forces. This means planetary orbits are mimicked. By carefully balancing forces which would alone lead to the explosion of the swarm against the gravity-like attractor forces and friction terms a slow and steady collapse can be enabled. As a consequence, the benefits of PSO are the controllable convergence behaviour and the smooth transition from large-scale scanning to small-scale local search.

The newly created hybrid EA described below incorporates many basic schemes of these popular EA paradigms. Its performance has been tested in comparison with a modern ES variant and two versions of PSO.

3.2 Description of the hybrid EA concept

3.2.1 The hybridisation concept (invention instead of choice)

If one has two foreign EA codes, and in order to decide which is the better one, they are both put to work on the same test function, then there is always some degree of uncertainty whether the results really tell something about the algorithm concept, or whether the performance difference is in reality caused by some other difference in the codes, e. g. differently acting mutation operators. A more trustworthy comparison can be made by rebuilding both EA codes on the basis of the same library of operators and subroutines. Another option, possible if the EAs are similar enough and offering yet more information gain, is to merge both algorithms into one single code containing a series of switches, so that by turning switch after switch the one EA is morphed step by step into the other one. If the EAs follow a generic (μ, λ) scheme of offspring generation, then soft fading is also possible instead of hard switching, simply by deciding separately for each new offspring chromosome through which algorithm scheme it will be generated. The decision can be made with random numbers and probabilities (e. g. blending 30 % GA offspring with 50 % DE and 20 % ES offspring) or with a fixed schedule (first generate 30 GA chromosomes, then 50 according to the DE scheme and 20 by ES). The great benefit is that one can not only compare the extreme cases but can also find out if the EA blends happen to be more efficient than the pure breed versions. If this is the case, one has quite straightforwardly invented a hybrid EA, and the blending ratios are knobs to tune its performance. This way, elements of ES, GA, and DE have been mixed and seasoned with a mutation cooldown schedule as in simulated annealing (SA) for controlling convergence. The result turned out to be performant on the chosen multimodal test problems.

A second idea going into this EA scheme is based on the observation that population bottlenecks are stressed in some biology schoolbooks as a feature helping in driving species away from old and towards new ecological niches and potentially supporting the formation of new species[3], i. e. they are pointed out as an explorative force. However, they do not seem to be a common feature of modern EAs. This was the inspiration for trying out a fractal scheme of selecting randomly from several populations and merging the selected chromosomes into one population to continue

[3]The role played by population bottlenecks in the evolution of species seems to be still under debate [295, 469]. Some more motivational thoughts will be added to the inspirational thoughts later on (section 3.3, p. 68).

with.

3.2.2 The generation cycle

For discussing the generation loop of the resulting algorithm, it is represented as pseudocode 2. Upon evaluation of the current population of parent chromosomes $\mathcal{P}_{\mathrm{F0}} = \{\vec{x}_i\}$ it is sorted according to fitness. The offspring population $\mathcal{P}_{\mathrm{F1}} = \{\vec{x}'_i\}$ is of same size and is divided into five segments or tiers:

$$\begin{aligned} \mathcal{P}_{\mathrm{F1}} &= \{\mathcal{P}_{\mathrm{elite}}, \mathcal{P}_{\mathrm{mutant}}, \mathcal{P}_{\mathrm{fp-mutant}}, \mathcal{P}_{\mathrm{CO}}, \mathcal{P}_{\mathrm{DE}}\} \\ &= \{\vec{x}'_1, .., \vec{x}'_{n_1}, \vec{x}'_{n_1+1}, \ldots, \vec{x}'_{n_4}, \vec{x}'_{n_4+1}, .., \vec{x}'_N\}, \end{aligned}$$

where $n_1 = N_{\mathrm{elite}}$, the size of the first tier, $n_2 = N_{\mathrm{elite}} + N_{\mathrm{mutant}}$ and so on. Each tier inherits genetic information from the sorted parent population through a different procedure in the following way: for all $\vec{x}'_i \in \mathcal{P}_{\mathrm{F1}}$:

- **if $\vec{x}'_i$ in 1st tier ("elite"):** Copy chromosome $\vec{x}_i$ and eventually mutate isotropically with strongly reduced step size. The purpose of this tier is elitism (i.e. conserving the best solution found so far) and local search around this solution.
- **if $\vec{x}'_i$ in 2nd tier ("mutants"):** Copy chromosome $\vec{x}_i$ and mutate. The purpose of this tier is to implement features of a (μ, λ)-ES.
- **if $\vec{x}'_i$ in 3rd tier ("free parent choice mutants"):** Choose a parent $\vec{x}_k$, copy it, and mutate. The purposes of this tier are selection pressure and gene pool diversity, it is one half of incorporating traditional GA-features.
- **if $\vec{x}'_i$ in 4th tier ("CO-bunch"):** Choose two numbers $k, l \in [1, N]$, $(k \neq l)$ and form the chromosome $\vec{x}'_i$ by mixing $\vec{x}_k$ with $\vec{x}_l$ using a GA-style crossing-over (CO) operator. Mutate with reduced step size. The purpose of this tier is to incorporate the other half of traditional GA-features.
- **if $\vec{x}'_i$ in 5th tier ("DE-bunch"):** Randomly choose three different parents $\vec{x}_k, \vec{x}_l, \vec{x}_m$ to treat them as in differential evolution (DE), i.e. add the scaled difference between two vectors to the third one. Mutate with reduced step size. The purpose of this tier is to let the population move in the search space like an amoeba by amplifying deviations of the population cloud's shape from sphericity.

The generation cycle is closed by evaluating the complete new set of chromosomes and renaming them parents. As the offspring is generated in tiers, and as these tiers represent common EA concepts, the above scheme could be called *tier-based hybrid EA* or short THEA.

From the different ways of selecting parent chromosomes, a fundamental difference arises between the first two tiers and the rest. That difference is also symbolically depicted in figure 3.1. In the first group the parent chromosome is copied directly into the next generation, that means $\vec{x}'_i$ is generated by copying $\vec{x}_i$ and then

Algorithm 2: THEA (tier-based hybrid EA)

$N \leftarrow 80$; $g \leftarrow 0$; $\gamma \leftarrow 0.04$; $\sigma \leftarrow 0.1$; $\vartheta_{c2u} \leftarrow 0.2$
$n_1, n_2, n_3, n_4 \leftarrow 4, 23, 42, 61$ `// tier boundaries`
$P_1, P_2 \leftarrow 1, 0.6$ `// mutation probabilities`
$\kappa_4, \kappa_5 \leftarrow \frac{1}{2}, \frac{1}{2}$ `// mutation damping`
$\mathcal{P}_{\mathrm{F0}} = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_N\} \leftarrow$ `GenerateRandomPopulation(`N`)` `// parents`
$\mathcal{P}_{\mathrm{F1}} = \{\vec{x}'_1, \vec{x}'_2, ..., \vec{x}'_N\} \leftarrow$ `GenerateEmptyPopulation(`N`)` `// offspring`
`EvaluateMembersOf(`$\mathcal{P}_{\mathrm{F0}}$`)`
`Sort(`$\mathcal{P}_{\mathrm{F0}}$`)`
while $g <= G$ **do**
  for $i \leftarrow 1$ **to** N **do**
    if $1 \leq i \leq n_1$ **then**
      $\kappa_1 \leftarrow$ `EliteMutationRate(`i`)` `// 1st tier`
      $\vec{x}'_i \leftarrow$ `Mutate(`$\vec{x}_i, P_1, \kappa_1\sigma$`)`
    else if $n_1 < i \leq n_2$ **then**
      $\vec{x}'_i \leftarrow$ `Mutate(`$\vec{x}_i, P_2, \sigma$`)` `// 2nd tier`
    else if $n_2 < i \leq n_3$ **then**
      $k \leftarrow$ `ParentSelect(`$N, 1, \mathit{pressure} = 1.0$`)` `// 3rd tier`
      $\vec{x}'_i \leftarrow$ `Mutate(`$\vec{x}_k, P_2, \sigma$`)`
    else if $n_3 < i \leq n_4$ **then**
      $k, l \leftarrow$ `ParentSelect(`$N, 2, \mathit{pressure} = 2.0$`)` `// 4th tier`
      if `Rand()` $< \vartheta_{c2u}$ **then**
        $\vec{x}'_i \leftarrow$ `CigarCrossover(`$\vec{x}_k, \vec{x}_l$`)`
      else
        $\vec{x}'_i \leftarrow$ `UniformCrossover(`$\vec{x}_k, \vec{x}_l$`)`
      $\vec{x}'_i \leftarrow$ `Mutate(`$\vec{x}'_i, P_2, \kappa_4\sigma$`)`
    else if $n_4 < i \leq N$ **then**
      $k \leftarrow$ `ParentSelect(`$N, 1, \mathit{pressure} = 4.0$`)` `// 5th tier`
      $l, m \leftarrow$ uniform selection from $\{1, ..., N\} \setminus k$
      $\vec{x}'_i \leftarrow$ `OffspringDE(`$\vec{x}_k, \vec{x}_l, \vec{x}_m$`)`
      $\vec{x}'_i \leftarrow$ `Mutate(`$\vec{x}'_i, P_2, \kappa_5\sigma$`)`
  $\sigma \leftarrow \sigma \cdot e^{-\gamma}$
  `EvaluateMembersOf(`$\mathcal{P}_{\mathrm{F1}}$`)`
  `Sort(`$\mathcal{P}_{\mathrm{F1}}$`)`
  $\mathcal{P}_{\mathrm{F0}} \leftarrow \mathcal{P}_{\mathrm{F1}}$ `// offspring become new parents`

applying the mutation operator with varying intensity. In the second group, however, the parent choice happens through random selection with varying degrees of selection pressure. The three tiers in that second group are distinguished mainly by the number of parent chromosomes being used to generate the offspring chromosome. For the 3rd, 4th and 5th tier this involves 1, 2 and 3 parents, respectively.

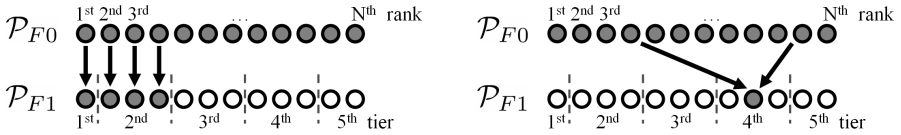


Figure 3.1 Copying versus parent selection
Symbolic sketch showing the two different ways of parent selection: direct copying and random choice under selection pressure. Direct copying is applied in the 1st and 2nd tier (shown on the left). For each new offspring chromosome in the 3rd, 4th, and 5th tier new random parent selections are made from the whole set of chromosomes in $\mathcal{P}_{\mathrm{F0}}$ (shown on the right). But these random selections occur with a certain amount of selection pressure, i. e. better ranked members of $\mathcal{P}_{\mathrm{F0}}$ are favoured over lower quality chromosomes. New chromosomes in the 3rd, 4th, and 5th tier are formed on the basis of 1, 2, and 3 parent chromosomes, respectively.

The following section explains the operators of the generational cycle and mentions details of the general architecture. The state parameter settings as chosen for the resonator optimisation will be given explicitly in section 3.4.

3.2.3 Operators and rules of the generation cycle

The **mutation operator** manipulates the different vector components of a chromosome $\vec{x}$ independently with probability P and consists of adding a random number δ of a standard normal distribution $\mathcal{N}(\mu{=}0, \sigma^2)$ with σ being the mutation step size parameter. σ is a scalar state variable of the algorithm, however, as it has been written for the general case of a bounded search domain where each component of $\vec{x}$ is allowed to cover a different interval $x_j \in [xmin_j, xmax_j]$, the δ-values are multiplied with the corresponding domain widths $w_j = xmax_j - xmin_j$, thus $x'_j = x_j + w_j\delta_j$. After initialisation, σ is multiplied with $e^{-\gamma}$ after each generation, a cool-down scheme inspired by simulated annealing (SA). In tiers 1, 4, and 5 the mutation step size parameter is scaled by the factors κ_1, κ_4, and κ_5. In the 1st tier, κ_1 is dependent on i and goes in equal steps from 0 to $\kappa_{1,\max} < 1$, which means the best member of $\mathcal{P}_{F0}$ is conserved untouched and mutation for the rest is damped. κ_4 and κ_5 are constants chosen from the interval $[0, 1]$ because higher values are deemed to create too much noise and marginalise the tier distinctions. The probability P that a gene x_j will be changed by the mutation operator is set to a value $0 < P \leq 1$ for all tiers except the first one, where it is $P = 1$. Next to information conservation the first tier's purpose is a local search in the direct vicinity of the current best chromosomes, and spatial isotropy is preferred here because the rationale is similar to evolution strategies.

Recombination operators are used in the 4th tier for mixing two chromosomes and in the 5th tier for mixing 3 parents. The crossing-over (CO) operators applied when mixing two parent chromosomes $\vec{x}_k, \vec{x}_l$ are *uniform CO* and a routine we call *cigar-CO*. *Uniform CO* means deciding independently for each gene from which parent (with equal probabilities) to take it. Note that this CO operator is

coordinate system-dependent, it can only create points in the corners of the coordinate system-aligned cuboid spanned by $\vec{x}_k$ and $\vec{x}_l$. The other CO operator, supposed as counterbalance, is not coordinate system-dependent. The idea is to create new sample points close to the line connecting the two parents. Now, $\vec{x}_k$ is assumed to be the better parent. It consists in creating a point $\vec{x}'_i = \vec{x}_k + r(\vec{x}_l - \vec{x}_k)$ ("extended line crossover") with r being a scalar random number uniformly sampled within the interval $[-\alpha, \beta]$ and a subsequent isotropic mutation (i.e. with probability $P = 1$ each vector entry x_j gets modified through addition of a random number from a standard normal distribution). The σ of this mutation is set to $a|\vec{x}_l - \vec{x}_k|$. Thus, this operator searches within and sometimes slightly outside a cigar-shaped volume between and a little beyond the two parents, and the aspect ratio a of that volume (not its thickness) is a control parameter of this CO operator. If $[-\alpha, \beta] = [0, 1]$, then the initial point (before mutation) is created only on the part of the line in between $\vec{x}_k$ and $\vec{x}_l$. Values of $\alpha > 0$ extend the searched line beyond the better parent, values of $\beta > 1$ push the limit beyond the worse parent.

The recombination routine mixing always three chromosomes $\vec{x}_k, \vec{x}_l, \vec{x}_m$ in the 5^{th} tier works almost like the traditional DE scheme [438]. By adding the scaled difference between two vectors to the third one, a new chromosome $\vec{x}'_i = \vec{x}_k + s(\vec{x}_m - \vec{x}_l)$ is created. The scaling factor s (called F in DE literature) is a uniformly distributed random number covering the interval $[0, s_{max}]$. The additional CO-step following at this point in DE tradition [438] is left out, as CO is already applied in the 4^{th} tier. Another justification for avoiding this particular CO operation is that it would water down the amount of information on the geometric shape of the parent population transported by the distribution of difference vectors.

The **selection pressure** is enacted through the parent selection routine f_{sel} applied in tiers 3 to 5. $f_{\text{sel}}(r, p)$ has to be based on pseudo-random numbers r and it needs to produce random integers $1 \leq \nu' \leq N$ such that lower values occur more often than higher ones implying a differential advantage for higher-ranked chromosomes to be selected more often for reproduction. p is thought to be the parameter for controlling the level of selection pressure, the tuning parameter for the steepness of the distribution. This is achieved by transforming the uniformly distributed random number $r \in [0, 1]$ in the following way:

$$\nu = \text{ceil}(\theta(r)), \quad r \in [0, 1],$$

with the transformation function θ defined as

$$\theta(r) = -\frac{N}{p} \ln(r), \quad \theta : [0, 1] \to [0, \infty].$$

N is the population size and p the selection pressure. The function θ transforms the uniform distribution of r into an exponential distribution. ν are integer numbers gained by rounding up the output of θ, thus $1 \leq \nu \leq \infty$. But the output ν' of f_{sel} should be restricted to $1 \leq \nu' \leq N$. A computationally efficient way of bringing numbers $\nu' > N$ back into the interval $[1, N]$ would be replacing ν' with modulo of ν', but such interval enforcement via projection has a relatively heavy distortion effect. A more neutral way is to try resampling if $\nu > N$. To limit resource use, the resampling iterations can be limited to an amount of l_{maxiter} iterations. Thereafter,

f_{sel} resorts to drawing a uniformly distributed random integer from $\{1, 2, ..., N\}$. Figure 3.2 illustrates relative selection probabilities produced by $f_{\text{sel}}(r, N, p, l_{\text{maxiter}} = 3)$ under variation of the pressure p.

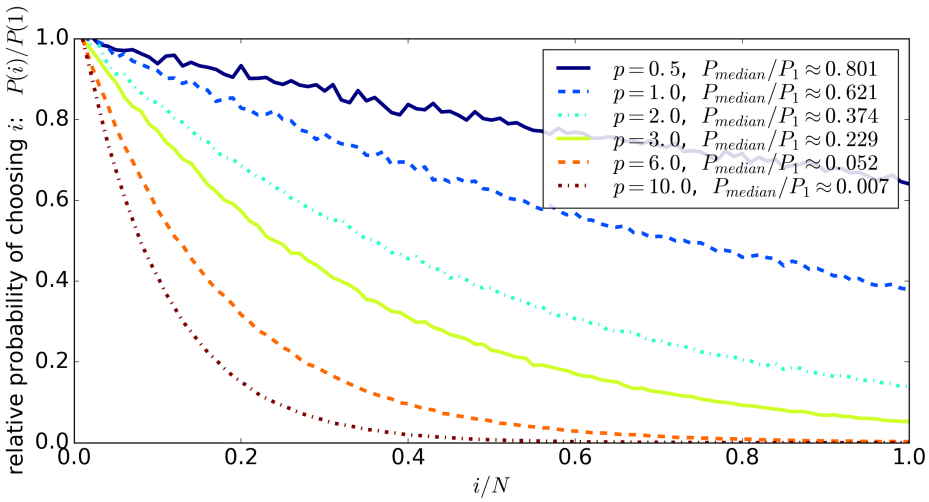


Figure 3.2 The selection pressure function
Statistics on the output of the parent selection routine $f_{\text{sel}}(r, N, p, l_{\text{maxiter}} = 3)$. The six data sets correspond to normalised histograms, that can be read as measured relative probabilities of choosing the i^{th} member of the parent population of size N. Relative probabilities were measured by normalising the frequency of choosing i with the frequency of choosing $i = 1$. Each set has been gained by 10000 calls to the function with $N = 100$ and the selection pressure p indicated in the legend. The legend also lists the measured relative probability of choosing the number of the parent with median fitness P_{med}/P_{best}.

Treatment of **search domain boundaries:** The problem of a sample vector created outside the bounded search domain is resolved by mirroring it back across the boundary, and should this point still lie outside, by filling new allowable (uniformly distributed) random values into those vector components that are out of bounds. The fitness evaluation comes afterwards.

3.2.4 The population merging scheme

A fractal-structured population merging scheme has been implemented with the aim of increasing robustness against local optima. The scheme is sketched in figure 3.3. Where the simplified sketch shows divisions into three subbranches, the used algorithm was implemented with a branching factor of $b = 4$. This results in the following scheme. Instead of initialising the population just once with random chromosomes, there are four consecutive generations of pure random chromosomes. Skimming the best quarter of trials from each one and collecting them in a new population results in a starting population which goes through the generation cycle for G_{proto} generations. Four such *proto-populations* are created and optimised alike, and by selecting 25% from the four final proto-populations, yet another starting population is formed. But this second condensation step is not performed by simply skimming the best quarter of each proto-population. Instead, the algorithm proved more robust when allowing also some randomly chosen chromosomes of lower quality and thus more gene pool diversity into the condensed population. Therefore,

after making sure that the single best individual of each proto-population has been copied, the rest of the choices occurs through applying $f_{\mathrm{sel}}(p)$ with a relatively low selection pressure. The mutation step size σ stays constant throughout evolving the four proto-populations, only while evolving the final population, it is made to shrink according to the above-mentioned cool-down scheme.

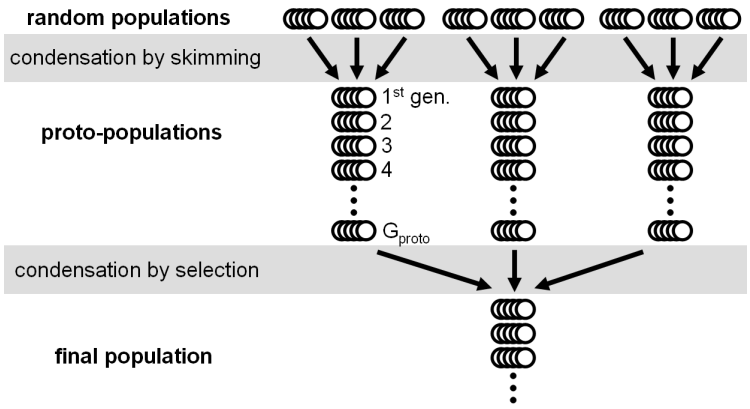


Figure 3.3 The population merging scheme
The best members of random populations are merged to form proto-populations, and high-fitness subsets of these are merged to form a condensed final population. For the sake of simplicity, the scheme is shown here with a branching factor $b = 3$, whereas in the EA implementation it is $b = 4$. The underlying motivation is the assumed usefulness of population dynamics (e.g. population bottlenecks) in the theory of biological evolution.

Fixed program versus self-adaptation: by intention, this hybrid EA was implemented without any features of self-adaptation. The assumed usage case is to stop the algorithm once the mutation step size has decayed to a reasonably small number and to continue with a dedicated local search algorithm from there.

3.3 Reasons behind this EA scheme

3.3.1 ES versus GA elements

In evolution strategies the focus is often placed on position, size, and shape of the population cloud and their change rates. The population cloud is a multivariate distribution and anchored at one place $\vec{x}_{anchor}$ in the genotype space. While concentrating on where to anchor the cloud next, i.e. deciding on which single chromosome to base the whole next generation, it is only consequent not to care about the conservation of gene pool diversity in the original biological sense. In ES the mutation operator recreates gene pool diversity in each generation, but combinations of particular gene values are always lost. By contrast, in canonical GAs the construction, inheritance, snippet recombination, and accumulation of beneficial gene combinations is all that counts. Such GAs are efficient on separable problems like the unrotated Rastrigin function, but they fail on many relevant nonseparable problems [224, 334, 394, 510, 511]. Relevant real-world problems not solvable by slimmer techniques than EAs are very likely nonseparable. However, it is also unlikely for many real-world problems (e.g. engineering problems) that each parameter has an

equally strong coupling with every other parameter. Imagine optimising an aeroplane design where five genes code for the wing shape and five other genes for the tail fin: here it makes sense to remember successful combinations of the one set while varying the other. Therefore it was tried to group ES-like with GA-like facets in the presented algorithm. The fact that a fixed-size fraction of high-quality chromosomes (the first two tiers) has an ensured reproduction probability of 1 is one adopted ES paradigm. Others are the use of a coordinate system-independent isotropic mutation operator and the fact the few best chromosomes will be reused very often. On the other hand, the use of *uniform CO* and the damped mutation applied to the individuals created by it, are GA-like features. Further GA-like aspects are the use of a mutation operator not affecting all x_j of a chromosome and the form of the parent selection distribution which also gives low-quality solutions a nonzero chance of reproduction. These facets are aiming at gene pool diversity and storage of valuable genetic information.

The tested CO operators of GA tradition also included BLX-α-β and Wright's heuristic CO (WHX) [202], but their use resulted in decreased efficiency while testing on the charged marble problem and Whitley's function F101 [508, 511].

3.3.2 DE, cigars, and more geometry thoughts

Assuming again that in many interesting EA application cases groups of parameters are more or less strongly coupled (which can be an argument for the use of traditional GA-like CO operators as *n-point-CO* and *uniform CO*), should one also assume that these correlations have particular relations with the coordinate axes? Probably not. This is the reasoning behind CMA-ES [196] where the multivariate mutation distribution is made to rotate completely independently of the genotype coordinate system. The resulting image is that CMA-ES makes the population cloud stretch and move towards neighbouring regions showing more favourable conditions like an amoeba does. In DE similar effects are produced without the existence of internal parameters responsible for cloud shape control and a corresponding update process: the mere addition of improved chromosomes will stretch the population cloud, and adding difference vectors found among random member pairs within the population to third member vectors adds an amplifying tendency to all deviations of the cloud shape from sphericity. This is why it was decided to add a tier using the DE scheme [438]. However, the *uniform CO* step, which commonly represents the last part of generating DE offspring, is left out for two reasons. First, CO operations take already place in other parts of the algorithm. Secondly, the uniform CO, because it spans up wide coordinate system-aligned cuboids, diminishes any coordinate system-independent effects of the other DE steps. Finally, the above-mentioned *cigar-CO* was added to have also one coordinate system-independent recombination operator for the case of two parent chromosomes. Furthermore, allowing values of $\alpha < 0$ with this operator offers a way to extrapolate gradient information given by parent pairs.

3.3.3 Search domain boundaries

Once sample vectors have been created which lie outside the bounded search domain, these boundaries can in principle be enforced by repeating any random-influenced

sampling routine until a suitable vector is found. But if such distribution sampling repetitions are to be avoided, the remaining simple options are (a) penalising the fitness, (b) pushing the vector back onto the boundary, (c) mirroring back, or (d) cycling back in from the other side. Possibility (a) was excluded in order not to waste computation time slots in a parallelised population evaluation scheme. The exclusion of option (b) is founded on the intention not to waste too many trials on a tiny fraction of the search space. The remaining decision between cycling (c) or mirroring (d) back is best made dependent on whether the objective function is periodic or not.

3.3.4 The population merging scheme

When from several populations only a small part is selected and these parts are merged to form a new population, then two effects take place. From the perspective of all the old populations this is a population bottleneck. On the other hand, this is the view focused on the newly formed population, it is a measure to create a population of increased gene pool diversity. Yet another perspective to view the population merging scheme is through the comparison with the concept of restarts in stochastic search algorithms.

Population bottlenecks as a decisive ingredient in the natural evolution of species was the inspirational thought. But in fact, the theory of population bottlenecks seems to be still in motion, the circumstances under which bottlenecks can accelerate evolution are still debated [295, 469]. The gene pool is a storage container for useful information shaped and distilled by the past. A population bottleneck has thus a negative connotation: the partially selection-driven and partially random-influenced loss of a large part of the stored information is primarily a loss; the smaller the gene pool, the less comprehensive and performant the data library from where to draw genetic code for recombining new chromosomes. The examination of model systems of population dynamics seems to tell that strong interaction of genes (e. g. blocking each other or switching on other genes), multi-gene-dependence of phenotype traits, and diploidy are required in order to allow circumstances under which single events or sequences of bottlenecks and subsequent phases of rapid population growth (*founder flush*[4]) can become beneficial for the advancement of evolution and speciation (i. e. separation into distinct species). The situation in optimisation with EAs (EAO) is much simpler, thus the justification for forced patterns of population dynamics has to be made up with independent thoughts.

An important factor is gene pool diversity. Like many other EAs THEA has a tendency to converge which can be understood in two ways, as a tendency of the scattered cloud of chromosomes to shrink in size and, because of the GA-like features of THEA, as a tendency of successful chromosome snippets to dominate the gene pool. Tendencies of exploration and exploitation have to be balanced in each EA, whereby in THEA the main controls are σ and the selection pressures. In that context there are two important motivations for the population merging scheme.

- The EA can be geared towards a sharper search and quicker convergence by

[4] A founder flush is the rapid expansion into uninhabited territory. But "uninhabited" does not necessarily mean "empty". It can also mean the expansion into a newly conquered ecological niche.

a faster decay of σ and increased selection pressure if the regular formation of new populations by mixing of old populations can reliably re-establish genetic diversity.

- Random mutations constantly dilute and destroy the information content of the genetic material. Population merging increases gene pool diversity without the perturbation of genetic code and without the deletion of genetic information through the infusion of pure random numbers.

Yet another perspective to view the population merging scheme is through the comparison with the concepts of *repeated local search (RLS)* or *iterated local search (ILS)*.[5] RLS means that a short running and quickly converging local search (LS) algorithm is restarted many times at random locations in the search space. Each restart means a total loss of memory. This means that from the first to the last each run has the same likelihood of being the one producing the particular solution counting as the best end result. ILS is not much different from RLS. A local search is started many times with a small budget. The only difference is that it is not restarted at a new random point each time, but in the neighbourhood of the final solution of the preceding run. A *perturbation operator* is applied to the best solution of the preceding search before the restart. The scattering range of the perturbation operator is the manifestation of the neighbourhood definition. Generally, the properties of the perturbation operator need not stay constant during the whole search. In ILS the first LS run has not the same likelihood of furnishing the best ever solution as the last run has. The likelihood is low in the beginning and increases steadily towards the last LS run with which the budget gets exhausted. The probability density of where the finding of the best ever solution occurs is shifted towards the end of the chain. Now comparing RLS and ILS with the population merging scheme of THEA, it is clear that the probability density for where the best ever solution occurs will be nonzero only in the final layers of the population tree, it will be shifted still more towards the end than in ILS.[6] Classifying RLS, ILS, and THEA by the amount of inter-population information flow, the situation looks like the following: in RLS there is now information flow at all. In ILS the information flow from one to the next population is in the form of one single chromosome handed over. This chromosome represents an ever larger search history, but its information content is diluted by the perturbation operator. In THEA the information input for a new population comes as N unperturbed chromosomes. All chromosomes of a newly formed population represent information gathered by preceding populations. Clearly the information

[5] By their names both RLS and ILS don't sound like relevant techniques for addressing global search problems, but in fact they are. This can immediately be seen by looking at two papers describing EAs that have turned out to be very performant in CEC competitions on real-parameter optimisation. One is the CMA-ES variant IPOP-CMA-ES by Auger & Hansen [17] implementing random restarts of the CMA-ES after stall detection; the only additional ingredient is a doubling of the population size after each restart. The other paper is the competition contribution of Liao & Stützle of 2013 [266]. They combine the very same IPOP-CMA-ES with an ILS scheme.

[6] In the presented setup of THEA, because of the first tier with the two purposes of conserving the best solution (elitism) and making the smallest local search steps, it is practically in 100 % of the cases that the best ever solution comes from the final population. But these thoughts about the population merging scheme are spelt out under the more general assumption that THEA may also be setup without any tier leading to elite-conserving behaviour.

of a chromosome skimmed from a pure random population is much less valuable than the information contained in one of the starting chromosomes of the last population. The gene pools handed over to each next layer in the population merging scheme represent the quintessence of exponentially growing search history trees. In summary it could be said that RLS can have the advantage of offering good global search properties with little conceptual overhead if a high number of restarts can be afforded so the scattering of starting points in the search space is not too sparse. As the distribution of the few starting points is so important in RLS, it is prone to being hampered by the curse of dimensionality. ILS connects the single runs of RLS by information flow. This can make it more powerful or less efficient in global search, the latter because ILS can only find valleys that can be reached from the one single starting point with the given amount of perturbation hops in the right directions. One single chromosome is a narrow bottleneck able to carry not a lot of information. In that context THEA can be seen as a scheme involving the piecewise evolution of populations in a way more suitable for global search. Information is handed over in the form of many diverse chromosomes combined to new gene pools. It does not have the disadvantage of relying on a continuous forward chain of perturbation jumps. Moreover, it avoids the disadvantage of periodic memory loss which would be a contradiction to the EA paradigm that the whole purpose of a global search run is the collection and leveraging of information gathered about the search landscape. In quite the opposite way it piles the memories of different pre-optimised populations on top of each other in a way increasing gene pool diversity, competition between pre-optimised solution alternatives, and the potential to recombine pre-optimised chromosomes of different types in useful ways.

The population merging scheme only has the consequence that one has to wait long until the most highly optimised chromosomes are likely to show up. This is no problem in a fixed-budget scenario where only the end result counts. But there are other thinkable scenarios where the number of consumed evaluation calls may be very variable and where a quick progress in solution quality is more important.[7] It may be the case in a real-world situation when the blackbox optimisation takes place under close human supervision and interference, when it is planned to stop, restart, retune, modify, or swap the algorithm manually each time it stalls and to leave it running as long as there is progress. This may be also the motivation behind a competition setup such as in [194] where not only the end result counts but also the budget it takes to reach a given solution quality. In that case the implementation of the population merging scheme might be seen as a drawback and single continuous runs of the THEA search could be the better way to go. One might think twice about the branching factor and the size of the population tree. Smaller population sizes are also considerable, in particular if the dimensionality of the search task is low. At least it should be realised that it matters in which sequence the populations are evolved. Figure 3.4 shows the two possibilities of evaluating the populations either layer by layer or branch by branch. The latter approach increases the likelihood of getting some of the highest-quality solutions earlier.

[7] That means the rationale behind the benchmarking goals described below on page 85 does not apply or is considered less relevant.

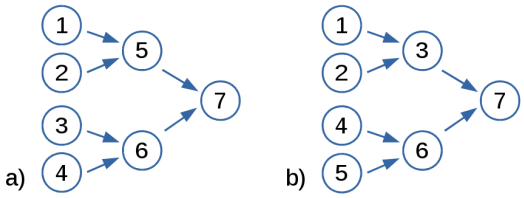


Figure 3.4 The question of sequence order in population merging
The fractal-like scheme of population merging allows for some flexibility in the way of sequencing the evaluations of the populations. Two possible sequence orders are shown: possibility (a) symbolises the evaluation by layers and (b) the evaluation by branch. The sequence order is irrelevant if only the end result counts, but it becomes relevant if importance is given to how early the algorithm comes up with its search results or if it is valued as beneficial when in some optimisation runs a demanded solution quality is achieved long before the exhaustion of the function call budget. In such a case variant (b) has an advantage over variant (a).

3.3.5 Why no adaptive features?

In this EA, the focus lies on the gene pool diversity of a large population, this means less available generations. A non-adaptive fixed cooling schedule ensures the population cloud has shrunken when the time is up.

Population clouds in most EAs sooner or later shrink. It is their purpose to zoom in on areas identified as promising. Watching evolution strategies acting on many common multimodal test problems, it often seems the option of expansion (through mutation step size increase) is implemented rather for symbolic reasons. If the shrinking process is determined by adaptive internal state variables of the algorithm, then the contraction may set in sooner or later depending on the random hits made during the first generations or it may be influenced by interference of the population cloud with the search domain boundary. For optimisation problems where dimensionality and complexity push up the lower limit for the population size and where a harsh upper limit for the amount of objective function calls comes from the computation cost a non-adaptive rigid EA concept can have a reliability advantage because there is no unknown delay for mechanisms of self-adaptation to reign in large populations. We implemented a rigid formula for the decay of σ inspired by the principle of simulated annealing (SA). Suboptimal time windows for population positioning dynamics are in this case the price to pay for reliably having a contracted population when the trials are used up. For the user this also means easy handling because of being relieved of trials and errors for tuning the adaptation procedure to indirectly achieve the goal of optimal contraction speeds fitting the application case, she or he just sets the speed which makes most sense when real-world boundary conditions may leave not much leeway anyway.

However, other parameters like the DE scaling factor s, the *cigar-CO*'s aspect ratio, the probability ratio between the two 2-parent recombination operators or probabilities for other momentarily excluded CO operators etc. can be imagined adaptive, but it has not been tried yet.

Should the tier sizes be made flexible? If it is assumed that the success of one tier follows also from good DNA snippets produced by another tier, then the first should not be made to grow at the cost of the latter, or at least not in the framework of a too simplistic feedback cycle.

THEA as EA experimentation lab

As already mentioned in the motivation of the algorithm architecture, the switches, knobs, and dials on this EA – like the tier boundaries, the CO operator blending ratios, the type and degree of damping of the mutation operators etc.– invite for experimentation. Its architecture makes THEA an experimentation lab for exploring the broad design space of merged basic EA schemes.

3.3.6 How the EA was tuned

Algorithm development, setup, and tuning is a diagnostics-driven feedback cycle. The shorter the time from trial to error or success, the better. The most careful regime of evaluating modified algorithm variants is based on sound statistics of many runs on a diverse set of test functions, but it has the severe disadvantage of not being the fastest regime. While the effect of slight changes in algorithm setup will often only be revealed by large statistics, fundamental changes in behaviour can immediately be spotted in diagrams of single search histories. During the design phase of THEA many quick checks and decisions were made analysing snapshot sequences of the optimiser running on specific test problems. The highest weight was put on the charged marble problem introduced below and further explained in appendix V.2. In both phases, the concept phase and the later phase of fine-tuning, score history plots as presented and described in figures 3.6-3.9 were used as a diagnostic tool. That approach of not entirely relying on statistics reduces the computational cost of EA tuning, is suitable for quickly evaluating drastic changes in the EA design and code, and thus in practice enlarges the design space being explored by the algorithm developer [428]. The disadvantage lies in the danger of losing objectivity. Therefore, performance evaluations by statistics should never be left out of the process. But, as the current task is to come up with an EA for solving a specific engineering problem and not with an EA exhibiting the best overall performance on a general set of functions with differing characteristics, a computationally slim and time-efficient manual human tuning process involving test problems deliberately chosen for their suitability in combination with punctually examining particular score history visualisations has been favoured.

The charged marble problem as a visualisable yet hard test function

In the context of predicting ideal crystal, molecule, or protein structures, Lennard-Jones problems are an important class of test problems for global optimisers and EAs [102, 306]. Minimising the potential energy of a number of repulsive particles like electrons which are constrained to move on a spherical surface is a strong abstraction which can be used as a simple test problem. The diagram in figure 3.5 on the left illustrates the two-dimensional case where the sphere is a ring. The task to distribute N particles evenly on a ring is too easy to be a meaningful EA test problem: any local search algorithm can reach the minimal energy by making the pattern equidistant; any rotation of the even pattern yields another optimal solution.

The charged marble problem was invented by trying to go out from this simple problem and make it harder by adding a second type of potential energy by

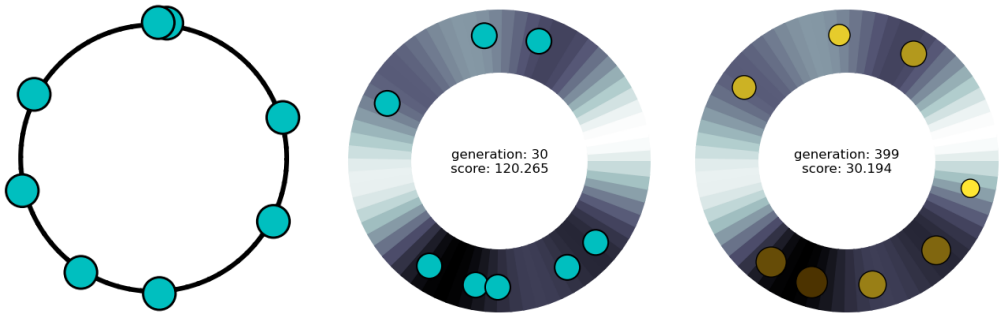


Figure 3.5 The charged marble problem as an optimisation test function
Going out from the easy task of minimising the total potential energy of a number of repulsive particles on a ring track (left), the charged marble problem was developed by adding a second type of potential energy: a track potential in the form of a smooth hilly landscape (centre). In the symbolism of the visualisation darker colours mean deeper valleys. As the four valleys are of different depth and separated by barriers of different height and width, it turns the distribution of the particles onto the valleys into a nontrivial task. In order to turn the nontrivial into a difficult optimisation task with unique local energy minima, another complication is added by associating the particles with different masses which can be symbolically expressed by giving the "marbles" individual colours and sizes (right). The difficulty level of particular problem instances can be controlled not only by the dimension, i. e. the number of particles, but also by the shape of the hill landscape.

associating the ring track with a hilly potential landscape. The middle image of figure 3.5 shows a visualisation of a candidate solution. The hilly track is symbolised by the background colour scale where a darker colour means lower energy. The depicted candidate is a partially optimised solution where the particle positions roughly match the valley regions. This test function has no unique global minimum as swapping particle pairs yields equivalent solutions. This makes it also feasible to find the global minimum by repeated local search initialised with random points with a low number of iterations: if the valleys receive the right numbers of particles in the initial random solution, then equilibrating the inter-particle potentials by incremental moves will quickly yield a near-optimal or the optimal solution.

After adding another complication in the form of giving the particles different masses (while keeping equal repulsive charges) the search for the global minimum becomes severely more difficult. The visualisation of this problem is depicted in the third image of figure 3.5. That the visualisation of the test problem remains easily interpretable and allows the distinction between good and bad solutions at a glance, and that moves of an algorithm in the search space can be directly associated with changes in the marble positions, these two properties can be of great advantage for optimisation algorithm developers who want to understand quickly how a new algorithm variant behaves.

The charged marble problem was published in a conference paper [428] together with A. G. Class and T. Schulenberg who accompanied the development of the test problem as discussion partners. The paper puts the charged marble problem in context by a literature review of visualisable test functions, whereas appendix V.2 is a shortened text on the main ideas.

Comparing THEA variants by score history diagrams

Figures 3.6 to 3.9 which are exemplarily highlighting the basis of some setup decisions show the scores of the whole chromosome population over time as coloured

scatter plots. They were gained from applications of THEA in various setups to the charged marble problem. Simple runs without population merging are shown. The populations are random-initialised (dark green dots in generation zero). The colour coding indicates the chromosome generation routine of each trial according to the legend given in the first figure. For example, the black-red subset of points represents the elite tier. The black dot in each generation is the first elite member which is conserved without mutation, and throughout the elite tier the brightness of the red colour correlates with both, the rank of the parent and the strength of the mutation operator. In the other three cases of tiers with non-uniform colours, the colour only represents the parent rank. For the DE tier it is the rank of that one of the three parent chromosomes to which the shift vector is added. As this parent is chosen with a strong selection pressure, the darker shadings are prevalent within the DE tier.

The two score history plots of figure 3.6 show the effect of annealing. In the upper plot the mutation step size reduction is turned on and γ is set to 0.05. The initial step size is $\sigma_0 = 0.08$. In the case represented below σ is kept constant at the value corresponding to the 25^{th} generation on the left. The run with constant σ achieved a final solution of only mediocre quality (judging by the histograms in figure 3.13; but of course, each plot pair has been deliberately selected for illustrative purposes) and no substantial improvements happen after generation 20. The interpretation is backed by plausibility thoughts: once a pre-optimised solution has been found, solutions which are yet better (i. e. even lower parts of the valley structure) might cover only a small fraction of the local neighbourhood of the pre-optimised solution. If larger mutation steps will simply jump across the deepest grooves of the valley structure, it is small steps which are needed.

Looking at single histories, any feature can of course be merely a consequence of randomness. But comparing two bunches of history plots from two setups allows for quick conclusions if relevant features are rare in one lot and frequent in the other. This way, the mutation step size reduction rate and the function call budget have been predetermined on the test problem before carrying the setup over to the resonator optimisation. Many other algorithm design and setup decisions have been made similarly. It is a way to insert a tight sequence of many objective (if caution is paid) decisions into the EA development process where there would have to be many guesses otherwise.

In the next plot pair in figure 3.7 the obvious difference is that the green dots scatter a lot more in the upper plot than in the lower one. This is due to a different setting of the sampling interval for the DE scaling factor s. In the upper case, it has been uniformly sampled from [0.6, 1.5], whereas the lower plot corresponds to [0.2, 0.6]. Large values of s mean that the DE tier becomes very explorative because offspring chromosomes are created farther away from the parents. The upper plot indicates an exaggeration of the explorative force of the DE tier because it produces solutions of much worse quality than all the other tiers. Secondly, the series of the best solutions at the bottom edge of the cloud, because it contains almost no green dots, shows that the DE tier almost never contributes directly to the search progress in that setup. The straightforward conclusion seems to be to favour the setting with lower values for s. But it is important to note that causality assumptions and

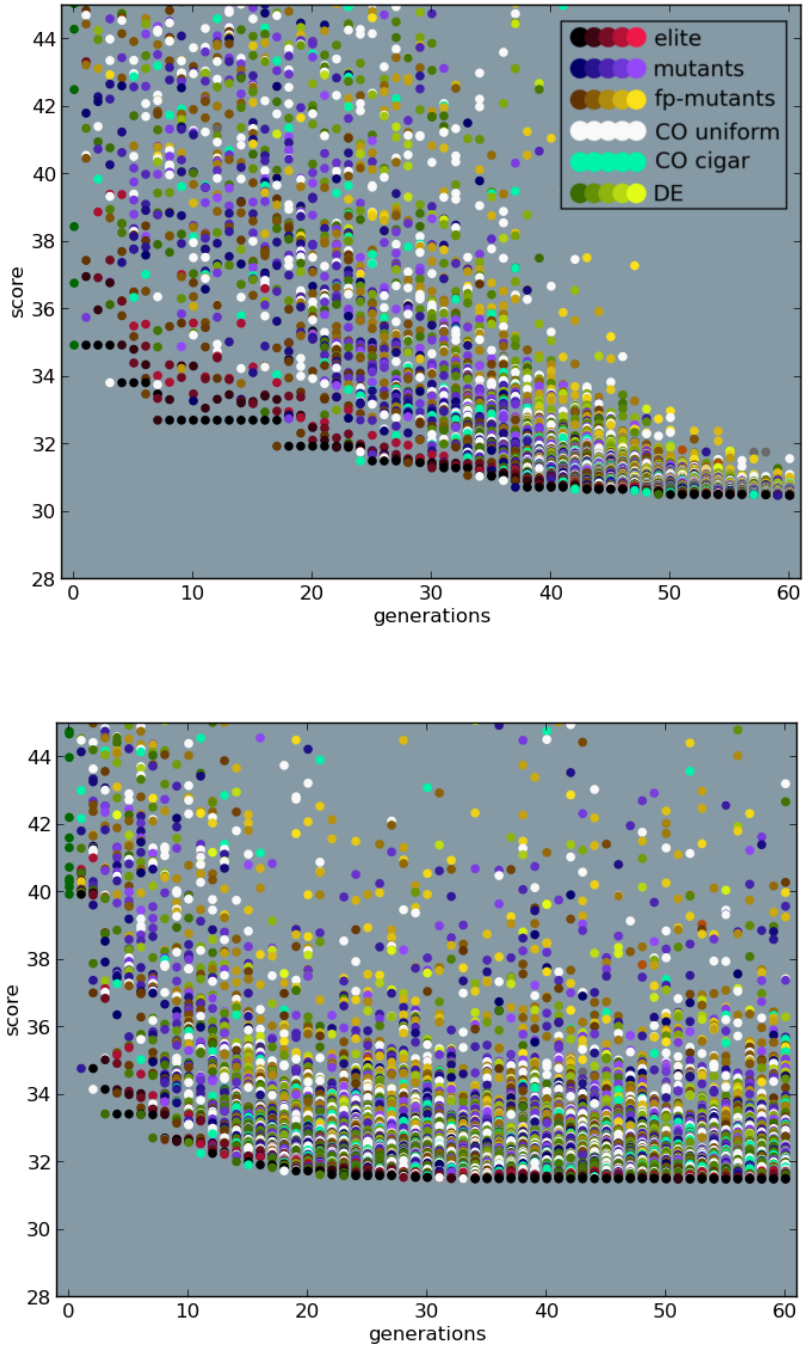


Figure 3.6 Visual diagnostics A: annealing
The two score history plots show the comparison of running THEA on the charged marble problem with exponentially decaying mutation step size (top) versus constant step size (bottom).

conclusions like this definitely need other support, at least by glancing also at the score of the final solution of a bunch of history plots, better yet by real statistics in such a fine-tuning question. The reason is that the tiers may contribute to the search progress indirectly which is explained in the context of figure 3.9.

The above two histories have been created with the exact same EA settings, in particular with σ initialised at $\sigma_0 = 0.08$ and $\gamma = 0.05$. These two examples simply show cases of convergence happening sooner or later in spite of the mutation step size decreasing along the same static pattern. For many EAs, convergence is coincidental with a narrowing of the chromosome population and a loss of gene pool diversity. Such is the case also for THEA. A scattered score cloud does not necessarily mean a bad thing as it was labelled in the discussion of figure 3.7. It is also a positive sign of active search before the chromosome population collapses. The plots of figure 3.8 are intended to illustrate the thoughts behind the decision in favour of a predetermined annealing schedule and against an adaptive step size control reacting to the search history: The annealing procedure still leaves the EA some freedom to collapse sooner or later, but it sets a time frame, it supports gene pool diversity in the early phase of the search and enforces the collapse in the later phase. It is responsible for generating the smaller search steps when they are needed when the search has to become more local. Most importantly, it prevents an early collapse, usually termed *premature convergence*, triggered by lucky random findings dominating the gene pool early on. This approach also reflects what's important in the real-world scenario of EA application: When many months are dedicated to tackling an engineering problem with optimisation algorithms, the computation time savings realised when a fraction of the optimisation runs will terminate a few hours earlier are of low relevance against the security that as much of the global search power as possible is used in each run.

The message of figure 3.8 can also be used to support the arguments made before in section 3.3.5 for increasing a search algorithm's reliability by avoiding unnecessary schemes of self-adaptation. Any self-adaptation scheme which can have the effect of a self-enforcing feedback cycle increases the dynamism which means that the searcher behaves differently from run to run.

The discussion of figure 3.8 touched the idea of adaptive algorithms looking at the aspect of mutation step size control. But an EA can be made adaptive in many other aspects of its state. The topics of adaptation or self-adaptation are abounding in EA literature. A common approach is to continuously keep testing several strategies in parallel and to increase the function call budget allocated to whatever has been more successful in the recent past. This could be very straightforwardly applied to the tier sizes in THEA: some success measures of the tiers could be taken as input signals and translated through some filtering into the tier sizes for the next generation. Figure 3.9 is intended to show that it is very easy to come up with ideas along these lines which would actually worsen the EA. This has to do with the fact that the tiers also contribute indirectly to the search progress. In THEA and in many other hybrid EAs each chromosome generation subroutine feeds on the genetic material produced by all the other subroutines. In the upper score history plotted in figure 3.9, it can be seen that white and cyan dots appear very often in the chain of best current solutions. That means the two segments of the two-parent CO tier (GA-

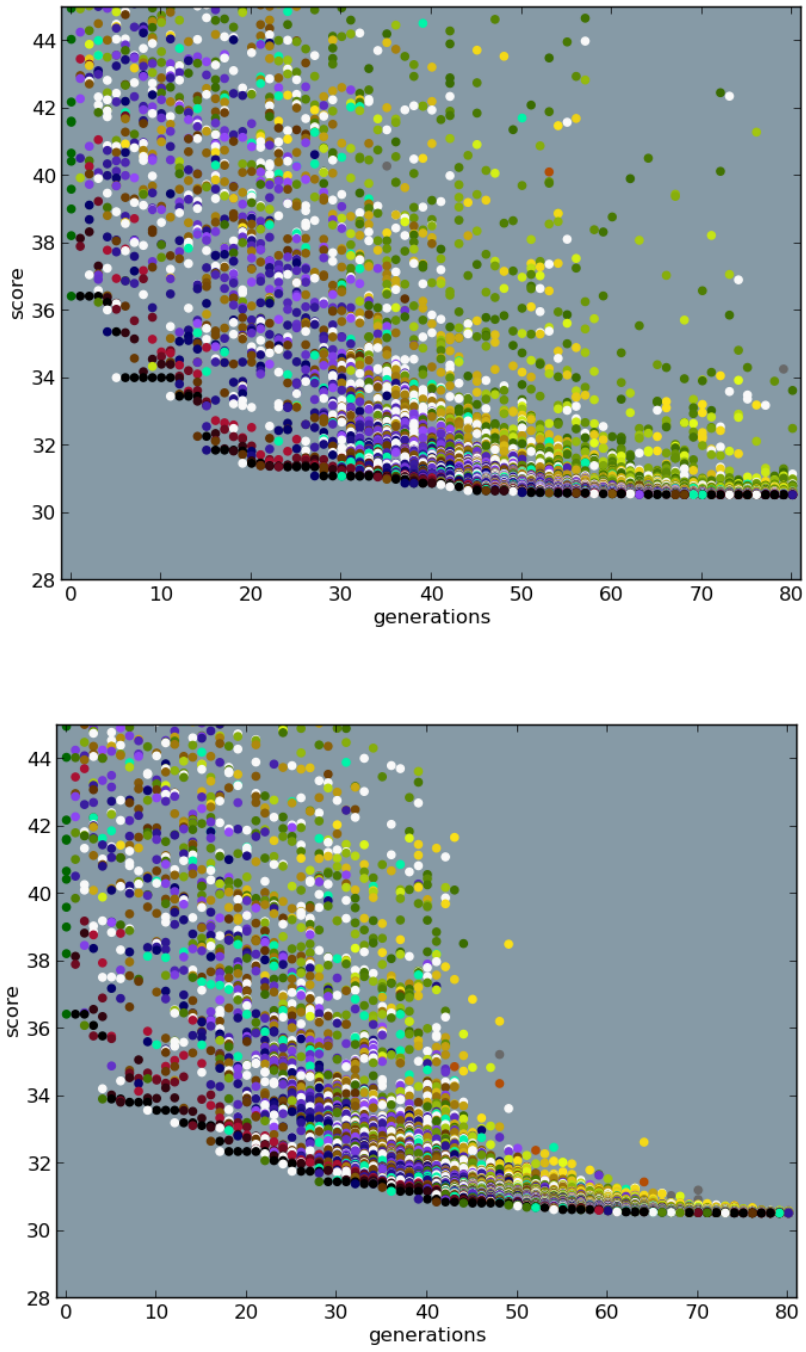


Figure 3.7 Visual diagnostics B: balancing tier cooperation
The difference in algorithm setting is the sampling interval for the DE scaling parameter s. In the EA run corresponding to the upper history plot it has been taken from the interval [0.6, 1.5], whereas the lower plot corresponds to [0.2, 0.6]. The larger values of s translate into a more explorative algorithm behaviour. The setup represented in the upper plot is much too explorative because the DE offspring in yellow-green yields candidate solutions of much lower quality than the other tiers and rarely contributes directly to improvements of the best solution.

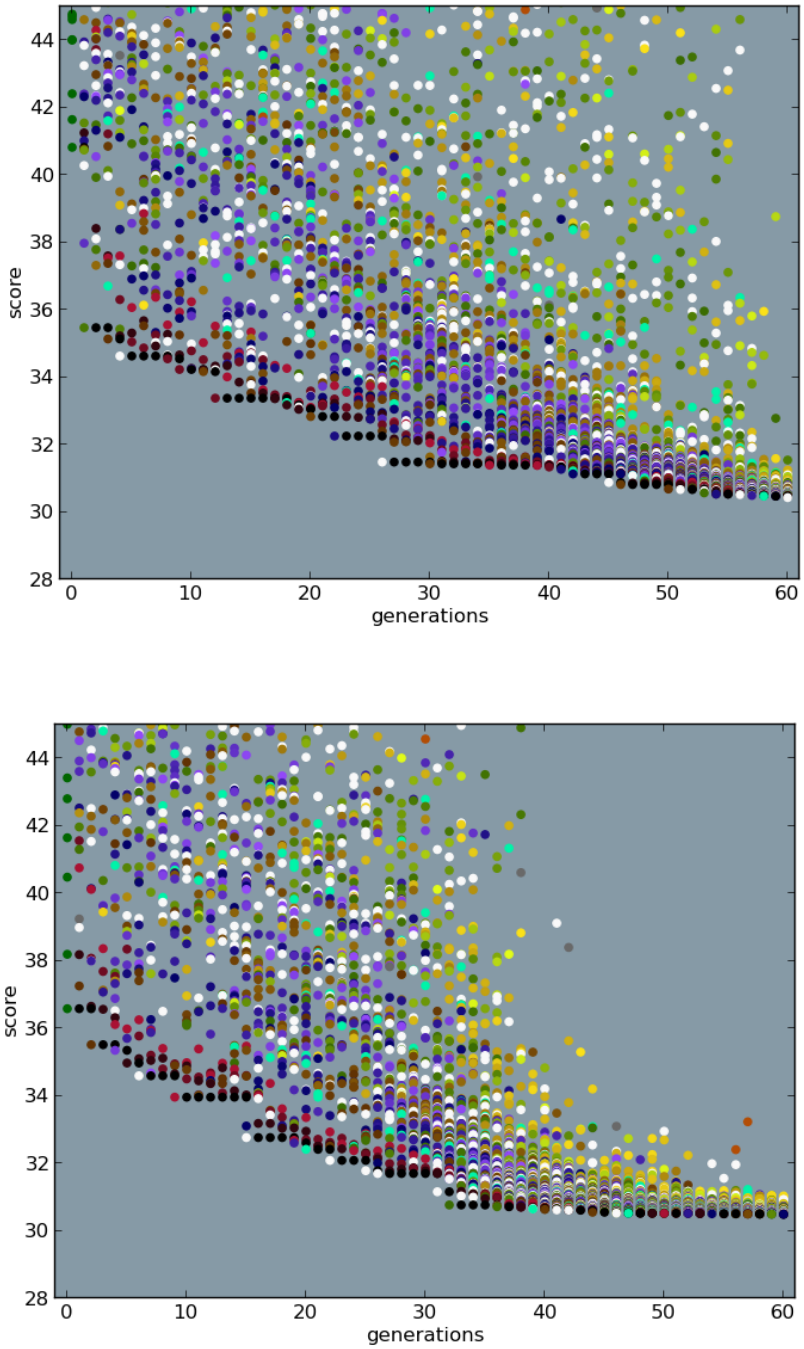


Figure 3.8 Visual diagnostics C: diversity collapse
The only difference between these two runs was the random seed. All algorithm settings were the same. The plots look so differently because the score spread collapsed much earlier in the lower plot as compared to the upper plot. The score spread implicitly reflects the gene pool diversity. For a user of stochastic search algorithms it is important to accept that one can never be in complete control of all search trajectories and outcomes. In the current context, the fact is taken to make an argument in favour of the annealing recipe, i. e. initially strong but steadily decaying forces boosting gene pool diversity, and against dynamic adaptation schemes.

style offspring) seem to be responsible for generating most of the trial chromosomes leading to optimisation progress. Should the GA tier thus be given more budget? The lower plot shows one history being representative of what happens more often when this is done and THEA is almost tuned into a conventional real-coded GA by dedicating 60 members of the population to the GA tier and only 5 to each other one. New information is then gained through the dampened isotropic mutation operator and the uniform CO ends up recombining snippets of DNA becoming ever more similar. This leads to more frequent stalls on mediocre solutions. The conclusion is that because the different parts of the hybrid EA work together, because each feeds on and contributes to the information stored in the gene pool, one should be cautious not to waste too much effort on too simplistic adaptation schemes.

Apart from that, if the objective function exhibits different properties on different length scales, then adaptation might always be too slow and provide for the currently experienced length scale a strategy mixture that would have been beneficial for searching the length scale just left but which is far from ideal for the moment.

Outlook: automated tuning

THEA with its variable tier sizes and independently tunable offspring generation routines, and not least its proven global search efficiency on particular multimodal search problems, seems to be a useful platform for a general-purpose EA. When intending to re-setup, extend, and tune it in the future for different and larger sets of test functions, statistics-driven automated tuning routines like "F-Race" [45] or "irace" [277, 278] should be considered as well. Algorithm tuning is its own optimisation problem.

3.3.7 Applied setup of the EA

The EA scheme and its implementation as Python code have been the target of constant experimentation and improvement. Note that during the history of resonator optimisations the EA setup was also further developed. During the latest resonator optimisations, the setup looked like this:

- **Population size and branching factor:** Four random populations were merged by skimming into one proto-population. Four proto-populations were condensed by random-influenced selections into the final EA population, thus $b = 4$. All populations were of $N = 80$ inhabitants.
- **Number of generations:** Proto-populations were evolved over 10 (sometimes up to 20) generations, the final population was kept evolving for at least 80 generations and continued for 40 or 80 more generations as long as substantial improvements to the best solution took place every couple of generations.
- **Mutation step size and annealing:** The mutation step size parameter σ has been initialised as $\sigma = 0.1$, the value from which it decays while evolving the final condensed population. After each generation the reduction happens according to $\sigma \leftarrow \sigma e^{-\gamma}$ with $\gamma = 0.04$. The mutation damping settings were $\kappa_{1,\max} = 0.2$ and $\kappa_4 = \kappa_5 = 0.5$. The mutation probability had been set to

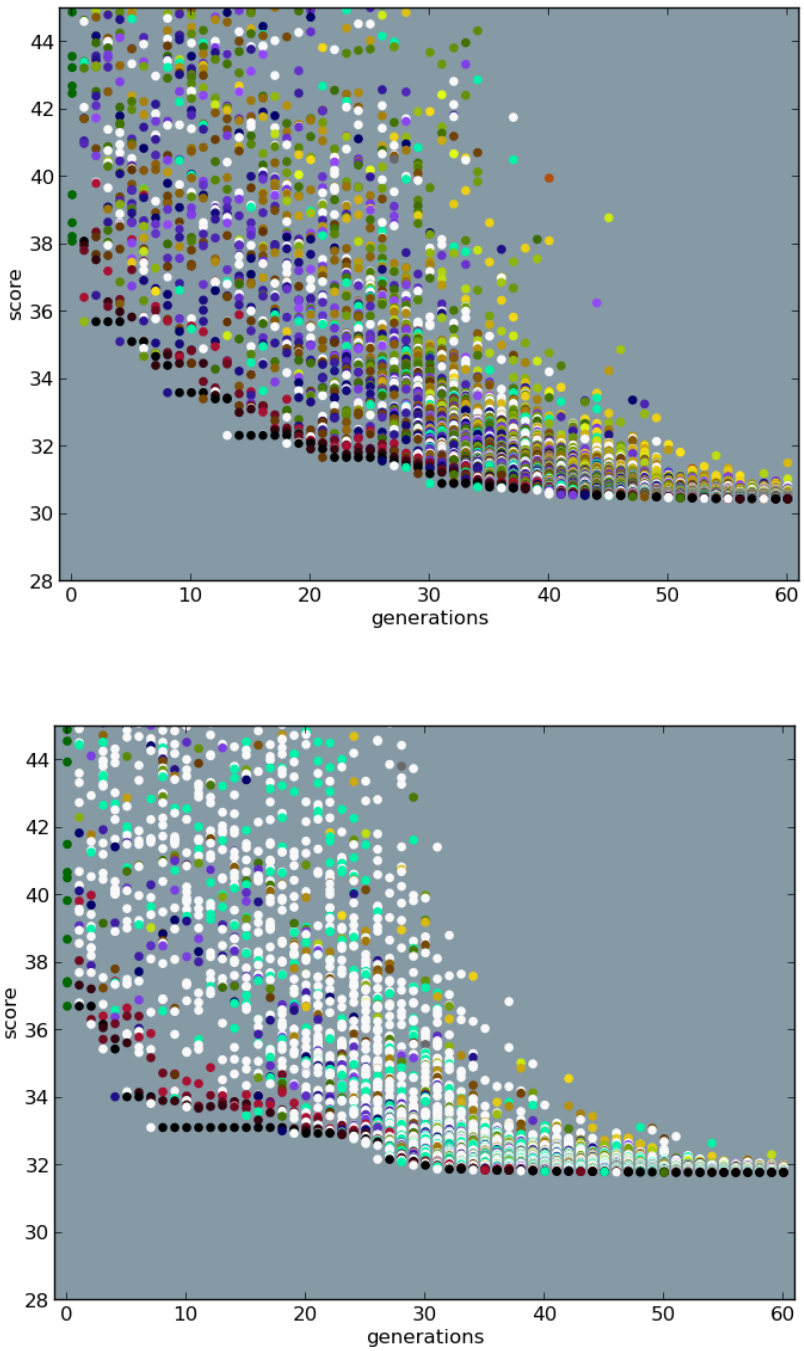


Figure 3.9 Visual diagnostics D: adaptation pitfalls
The upper score history shows by frequent white and cyan dots on the lower edge that GA-style CO offspring very often leads to improvements. The conclusion that a larger GA tier at the cost of all other tiers will improve the search performance is however wrong, and it can be easily proven. The lower diagram shows an exemplary history resulting from a setting with 75 % GA-style offspring. It exhibits a low frequency of improvement steps and a mediocre final solution. In cases where the property changes are drastic, obvious, and persistent, it suffices to collect a low number of history plots to make objective EA development decisions, e.g. against the attempt to implement a simplistic tier size adaptation scheme.

$P = 1$ during the latest optimisation runs. Before, it had been generally set to $P = 0.6$. Throughout the evolution of the proto-populations, σ is dampened by an additional factor of 0.3 but constant.

- **Tier sizes:** The sizes of the subdivisions of the offspring population corresponding to the different procedures of offspring generation were set to $[N_{\text{elite}}, N_{\text{mutant}}, N_{\text{fp-mutant}}, N_{\text{CO}}, N_{\text{DE}}] = [4, 19, 19, 19, 19]$.

- **Selection pressure:** For the three tiers $\mathcal{P}_{\text{fp-mutant}}$, $\mathcal{P}_{\text{CO}}$, and $\mathcal{P}_{\text{DE}}$ the selection pressure p had been set to 1, 2, and 4, respectively. For condensing the proto-populations into the final population, $p = 2.2$ was applied.

- **CO operators:** The ratio of cigar-CO versus uniform CO has been set to $\vartheta_{c2u} = 0.2$. The cigar-CO operator's aspect ratio was 10, and its $[-\alpha, \beta]$ setting (the bias for offspring creation near or beyond the parents) $[-0.8, 0.3]$, so that the cigar extends 0.8 times the inter-parent distance beyond the better parent and 0.3 beyond the less fit.

- **DE offspring:** The scaling factor s was sampled from the interval $[0.2, 0.8]$. This means avoiding the smallest mutation steps, which can be seen as a measure to support the exploration tendencies of that tier (similar as in [51]). Suppressing the local search tendencies with this measure is not deemed to be a major problem because other tiers are dedicated to local search.

This means that one proto-population starts with $4 \cdot 80 = 320$ random trials and costs $320 + 10 \cdot 80 = 1120$ functions evaluations in total. One whole algorithm run comprising four proto-populations and 80 generations of the final condensed population requires thus $4 \cdot 1120 + 6400 = 10880$ solution candidate evaluations and 200-500 more with a subsequent local search for solution fine-tuning using the downhill-simplex algorithm by Nelder & Mead [323].

Short descriptions of THEA as a new type of hybrid EA were published in two conference papers [429, 434] together with A. G. Class, T. Schulenberg, and R. T. Lahey Jr., who are proponents of sonofusion-related research at RPI and KIT. They accompanied the author's EA development and benchmarking efforts as advisory discussion partners.

3.4 Benchmarking the search algorithm

3.4.1 The test problems

In the setup just described, THEA was benchmarked against two state-of-the-art EAs, CMA-ES and PSO, by measuring the performance on five different test functions:

- the shifted rotated Weierstrass function from the CEC-2005 collection [444],

- the shifted expanded Rosenbrock and Griewank function from the CEC-2005 collection [444],

- the FM-synthesis problem from the CEC-2011 collection [102],
- the test function F101 by Whitley et al. [508, 511], and
- the charged marble problem [428].

Figure 3.10 shows illustrations of the five test problems. The FM-synthesis problem is defined as a six-dimensional search problem and the charged marble problem was instantiated as eight-dimensional function. The other three functions were all used in a ten-dimensional setup for this benchmark. Detailed descriptions of the test functions are given in appendix V. PSO and CMA-ES were used in different setups and code versions. By running each EA 400 times on each problem, large enough statistics were created to allow judgements not only based on means and medians but rather on clearly distinguishable differences in the shapes of solution quality distributions (figures 3.11-3.15, pp. 87ff).

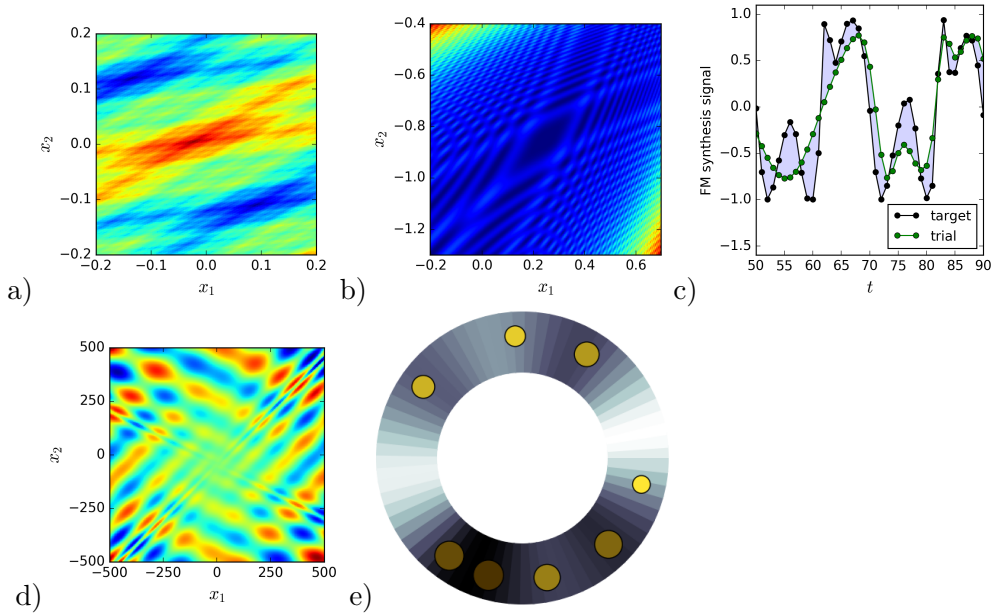


Figure 3.10 Illustrations of the five test problems
Plot (a) is a 2D instance of the shifted rotated Weierstrass function. From the orientation of the structure it can be seen that the used rotation matrix involved stretching and squeezing along different directions. (b) shows the shifted expanded Rosenbrock and Griewank function. Diagram (c) illustrates the FM synthesis optimisation task where a signal generated by a formula of nested sine functions has to be matched to a target signal. Plot (d) shows the function F101 where the global minimum is far away from the centre position and hidden behind high mountains. Diagram (e) illustrates the charged marble problem. In all 2D colour map plots the common "jet" colour scale is used where blue/red stands for low/high function values.

3.4.2 The competing EAs

THEA is being compared with CMA-ES (evolution strategy with covariance matrix adaptation, [196]) in three versions and PSO (particle swarm optimisation, [120, 230]) in two versions:

- CMA-ES-A: An own code implementation (Python with NumPy) based on Nikolaus Hansen's barebone Python code `barecmaes.py` [191] of early 2011.

It is run with a fixed population size as (40,80)-ES. This one as well as the other two CMA-ES versions are started with a mutation step size of $\sigma_0 = \frac{1}{3}$ times the search domain width. The initial anchoring is at a random location.

- CMA-ES-B: (40,80)-ES using Hansen's extensive production code version [191] of CMA-ES in Python (version of 2014).
- CMA-ES-C: Same code as above but in the setup with restarts ("IPOP-CMA-ES") as used by Auger & Hansen for participation in the CEC-2005 competition and described in [17]. The initial population size of the (μ,λ)-ES with $\mu = \lambda/2$ is $\lambda = 10$. The stopping criterion for triggering restarts is `tolfun`=0.1, and the population size is doubled after each restart. The restart involves a new random starting location and the resetting of the mutation distribution to the isotropic case with the wide initial σ_0, i. e. a total loss of memory.
- PSO-A: the PSO implementation in Python by Marcel Caraciolo [72], called `pyPSO`, executed in the "constricted" mode with[8] $c_1 = c_2 = 2.05$ and a population size of 80.
- PSO-B: An own PSO implementation [430] with[9] $\alpha = 0.7298$ and $\psi = 2.9922$. Transgressions of the search domain boundaries are treated by reflecting particle position and velocity back in. The swarm has a size of 80 and its communication network is a simple 8×10 von Neumann topology. Information is passed up to a local neighbourhood degree of 2.

All these EAs are based on open source codes.

3.4.3 Benchmarking goals and guidelines

The central goal of this benchmark is to identify the most suitable EA for applying it to an engineering problem where one evaluation corresponds to the simulation of a physical system, in particular the SF resonator FE simulation. Thousands of evaluations of this FE model are still affordable, whereas 10^5 simulations per optimisation run are impractical. Therefore, random-initialised optimisations with fixed budgets of 10 000 evaluation calls were the basis of the benchmarking. For 10D problems this is ten times less than the budget during recent CEC competitions [265, 444]. It has already been pointed out (see figure 3.8) that time savings[10] by EAs not needing the whole budget are a lesser concern than the security that in each run the full global search power is leveraged. So, in spite of alternative ways of judging the performance of optimisation algorithms, ways taking also account of

[8] Explanations for the modes and state variables of the PSO algorithm can be found in appendix T.4.7 where the PSO concept is outlined.

[9] dito

[10] The time scale of EA optimisation runs (days in the SF resonator case) is not the only one of relevance because the time scale of a person setting up a simulation, the optimising algorithm, and the interface between the two, is another crucial one. Savings on the small scale (e. g. an EA converging somewhat quicker every other run) are only beneficial if they are not likely to be connected with more necessary thinking and checking on the upper work layer of human trial and analysis (e. g. for finding out whether quick convergence means luck or wrong setup).

how quickly results come in [194], here, only the end results were considered, i. e. only the quality of the best solution ever found during one run was taken as the measure of success.

This competition setup with relatively low call budgets induces the problem that the final solution quality depends a lot on how much the EA has converged. At a moment when the genotype population has converged to a tiny size in a smooth search space, the EA is merely doing local search. However, dealing with costly engineering problems with a smooth objective function, local search (LS) can be done very efficiently with dedicated and slim algorithms like Solis-Wets [424], (1+1)-ES, or the deterministic downhill-simplex search (Nelder-Mead algorithm [323]). The EA should not waste its budget on local search and not be judged for its performance in this discipline. An EA which has found one peripheric part of the valley containing the global minimum should not be judged worse than a competing EA having reached closer proximity to the global minimum. Or, thinking of the curse of dimensionality and what it means for the cost of mapping any search space region, if two algorithms have found the valley containing the global minimum (or just the same valley), no difference in performance judgement should arise from how much luck may have been involved in possibly just few probes of that valley. The problem can be overcome and the benchmarking experiment made insensitive to the local search efficiencies of the EAs by finishing off each EA search with the same local search. This has been done by allocating the last 400 function evaluations to a downhill-simplex search starting with the best chromosome encountered so far. The simplex has then been initialised with an edge length of $\frac{1}{800}$ of the search domain width.[11]

The solution qualities after 400 such EA+LS optimisation runs can be nicely compared if their distributions are plotted as histograms like in figures 3.11 to 3.15. But how can winning be defined interpreting these histograms? The question has to be answered under consideration of the context. This means that, firstly, very thin tails of singular lucky best results have to be ignored because the EA cannot be run many times on the resonator optimisation and luck cannot be counted on. Next, the right halves of all the distributions can be ignored as well. New procedures need practice, thus there will be more than one EAO[12] run with the costly engineering problem of interest. How the worse half of the result distribution looks like is unimportant as soon as one assumes that probably not every trial will be compromised by bad luck. Consequently, the features of interest in the benchmarking histograms have to do with the left halves of the distributions and the left slopes, in particular where the left flank is positioned on the x-axis, how steep its slope is, and how narrowly packed the left part of the distribution is. With a focus on these points, the histograms will be discussed in two groups, first the two problems selected from the CEC-2005 benchmarking suite, then the three harder ones.

[11] Why such a small initial size? Looking up any description of the Nelder-Mead algorithm it can be quickly inferred that shrinking is much more expensive than expansion. However, a small initialisation size requires checking that the LS is not blocked by "solver noise", i. e. ruggedness or structure on f_{obj} arising from numerical computation methods like the FE method.

[12] Here and below *EAO* will be used as a shorthand for *optimisation with evolutionary algorithms* or *evolutionary algorithm optimisation*.

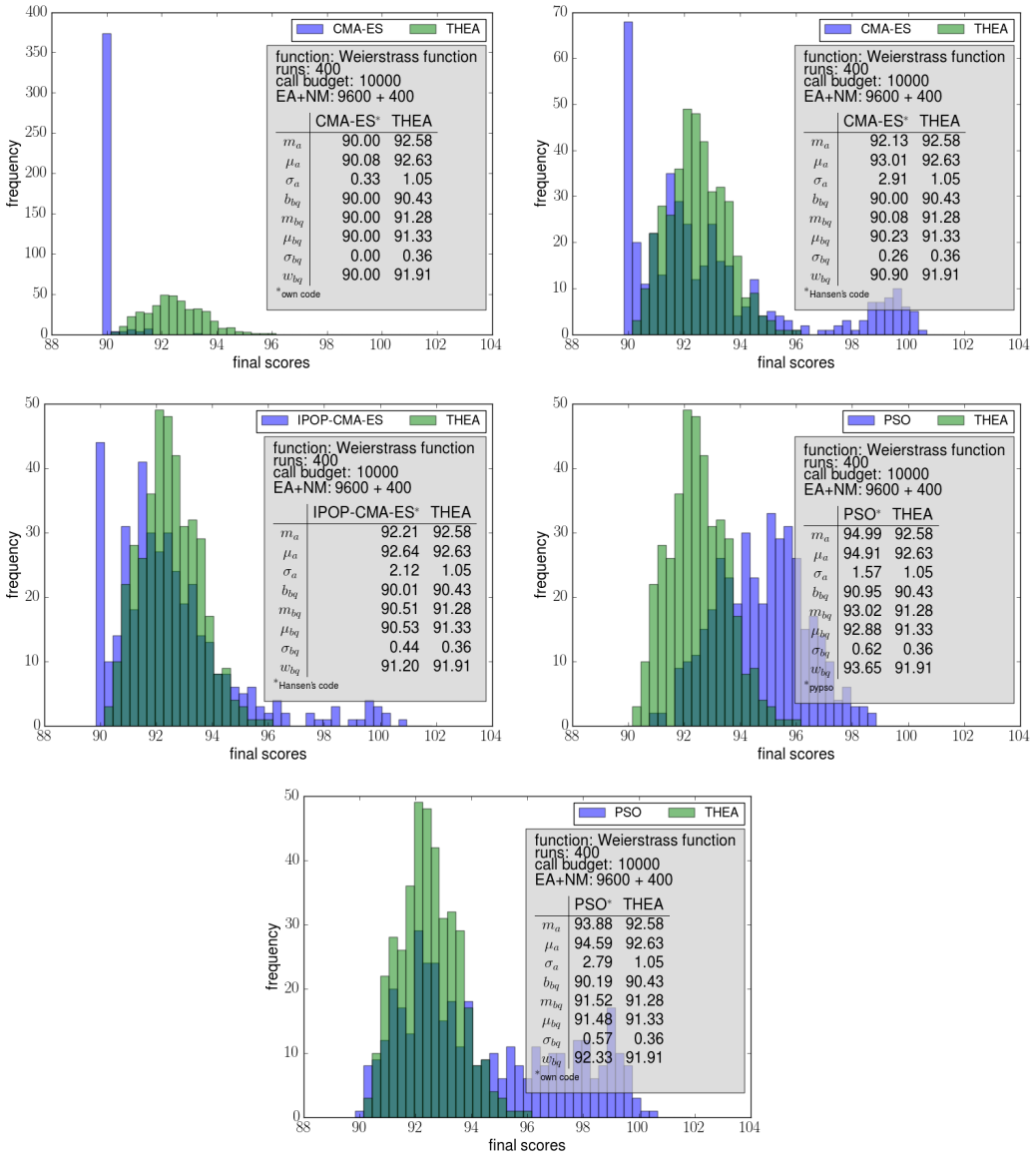


Figure 3.11 Benchmarking THEA on the Weierstrass function
All three variants of CMA-ES are clearly better than the competition in the search landscape spanned by the shifted rotated Weierstrass function [444]. The landscape is made up of a sum of twenty cosine functions per dimension covering an extremely wide frequency band, with amplitudes being inversely proportional to the spatial frequency. This allows the features of CMA-ES to play out their full potential of using the population cloud as a low-pass probe for the envelope of the valley bottoms (a property inferred from figure T.17, p. 540). As the mutation step size σ decays and the genotype population contracts to ever smaller length scales, the evolution strategy works itself through the cosines, from the coarser to the finer structures and follows the perceived averaged slopes (see figure T.16, p. 540).

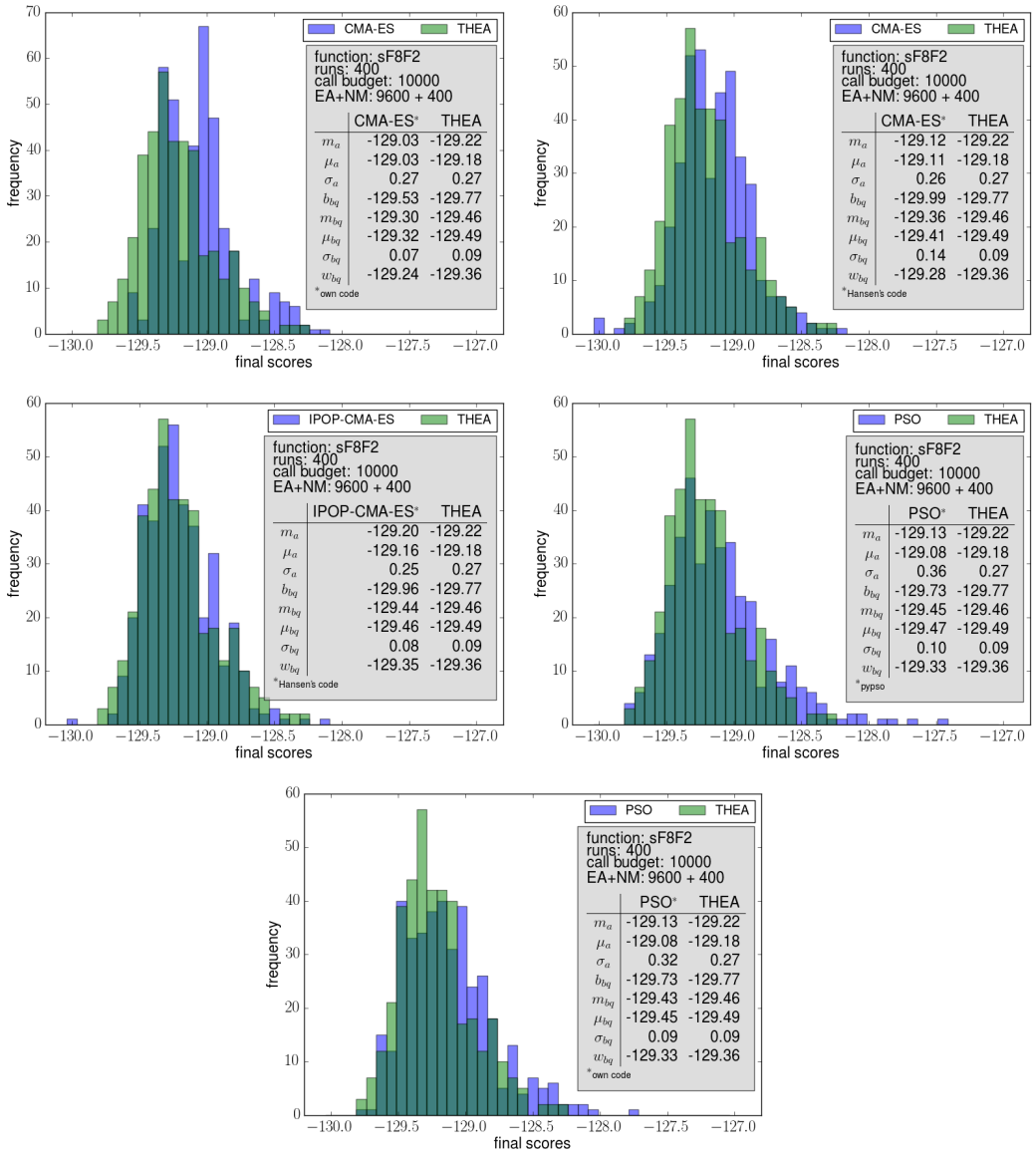


Figure 3.12 Benchmarking THEA on the eF8F2 function
On this test function the score distributions from the three algorithms match quite closely. In particular, all are of the same shape. Only CMA-ES-A falls behind because of being shifted substantially to the right. CMA-ES-B is also shifted a tiny bit to the right in comparison with THEA, but one should not overlook the leftmost blue bar in two of the plots indicating that only CMA-ES is able to find the global minimum sometimes. Therefore, this test function allows no declaration of winners or losers.

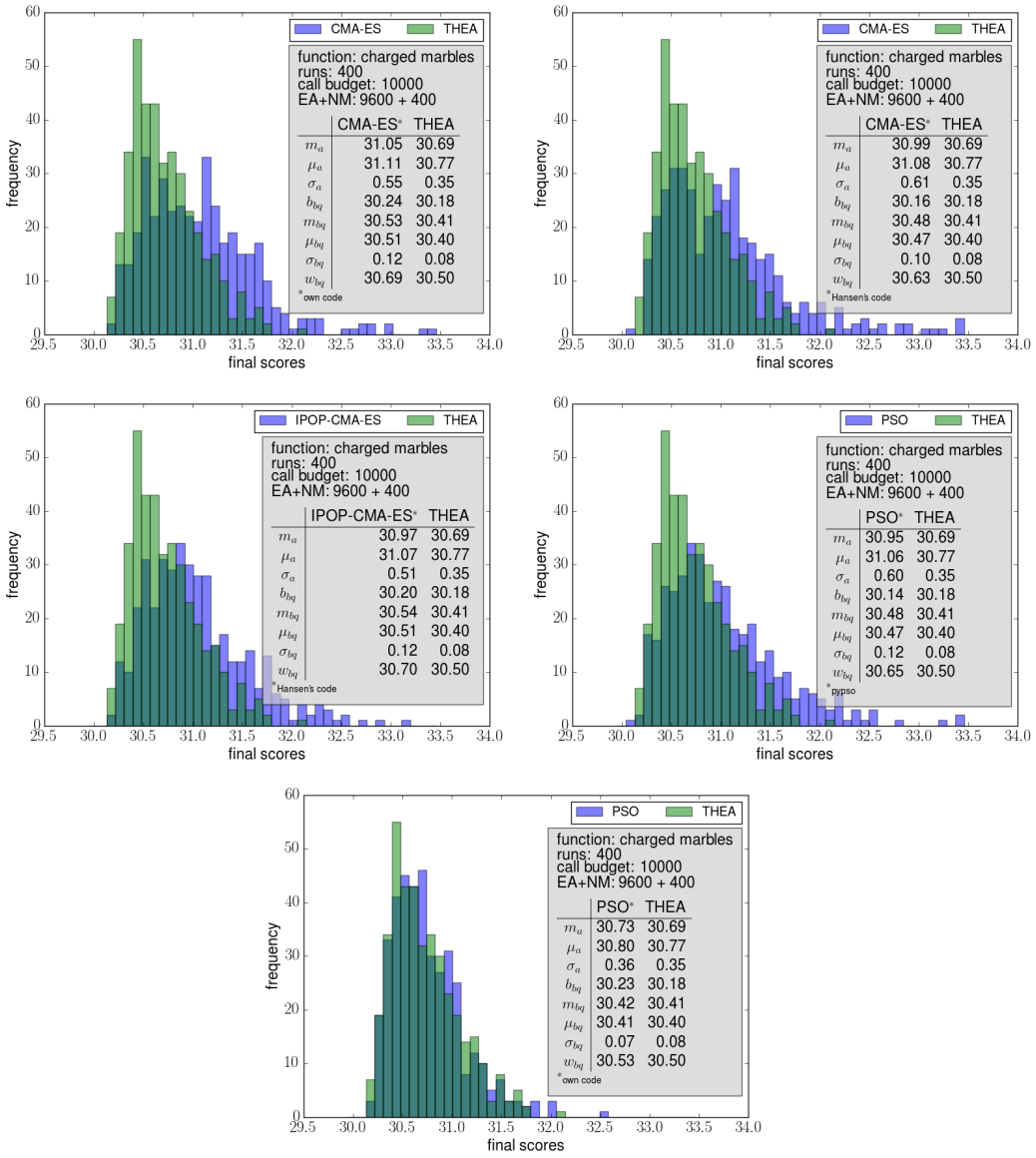


Figure 3.13 Benchmarking THEA on the charged marble problem
On this test problem THEA performs better than the other algorithms. In each comparison, the THEA histogram has its peak to the left of the competitor's peak, the left slope farther to the left, and, most importantly, the steepest left slope. Therefore, THEA is the EA promising the best global search performance and least wasteful usage of computational resources for those types of real-world optimisation problems where the search space characteristics are similar to the charged marble problem.

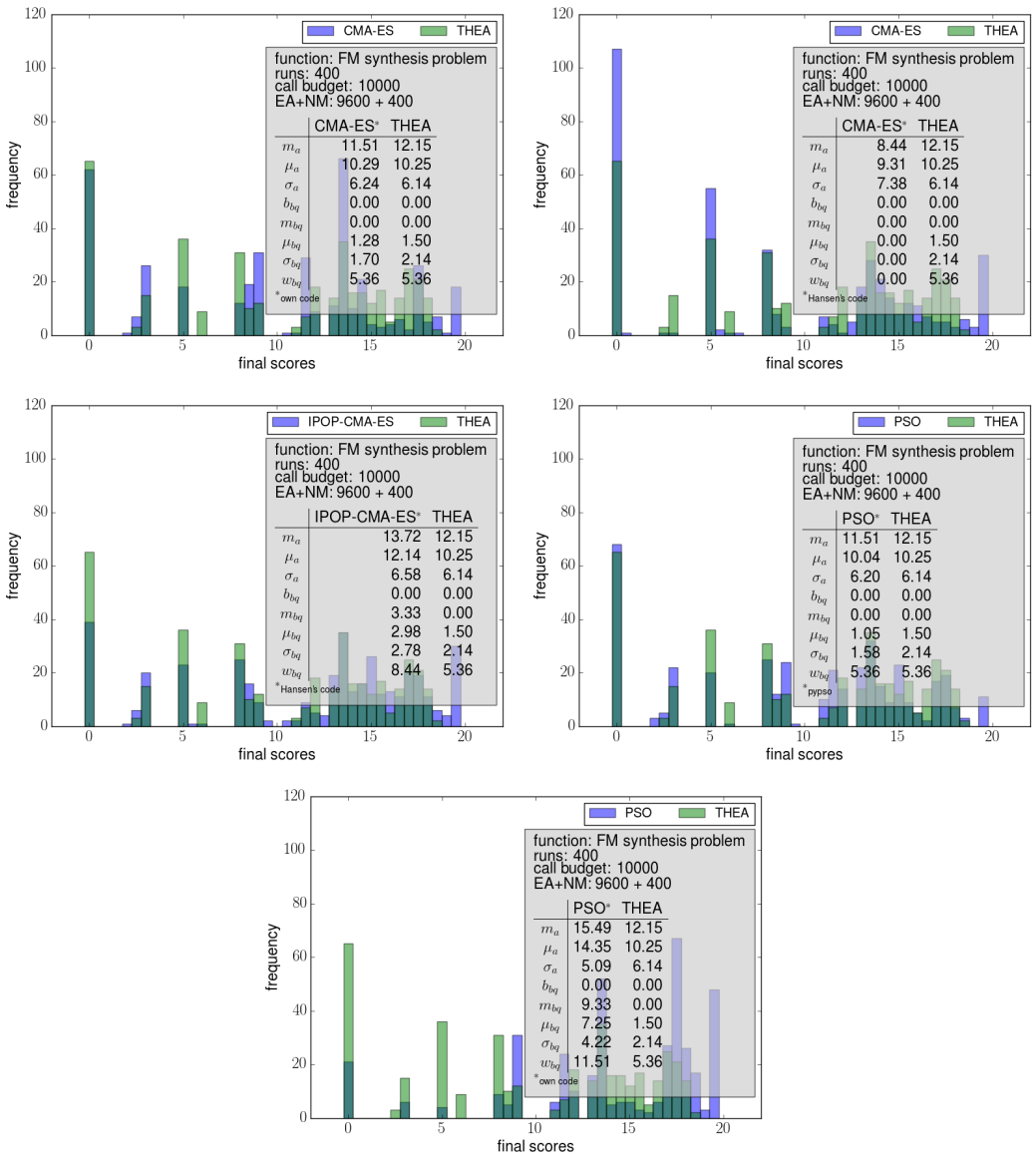


Figure 3.14 Benchmarking THEA on the FM-synthesis problem
These histograms reveal that the globally optimal and a few near-optimal solutions form discrete levels or bands, so the histograms are broken up. Only the solutions of inferior quality with $f_{\text{obj}} > 11$ form a bulk. All EAs find the optimal solution sometimes, but CMA-ES in the second setup stands out with the by far highest rate of finding it.

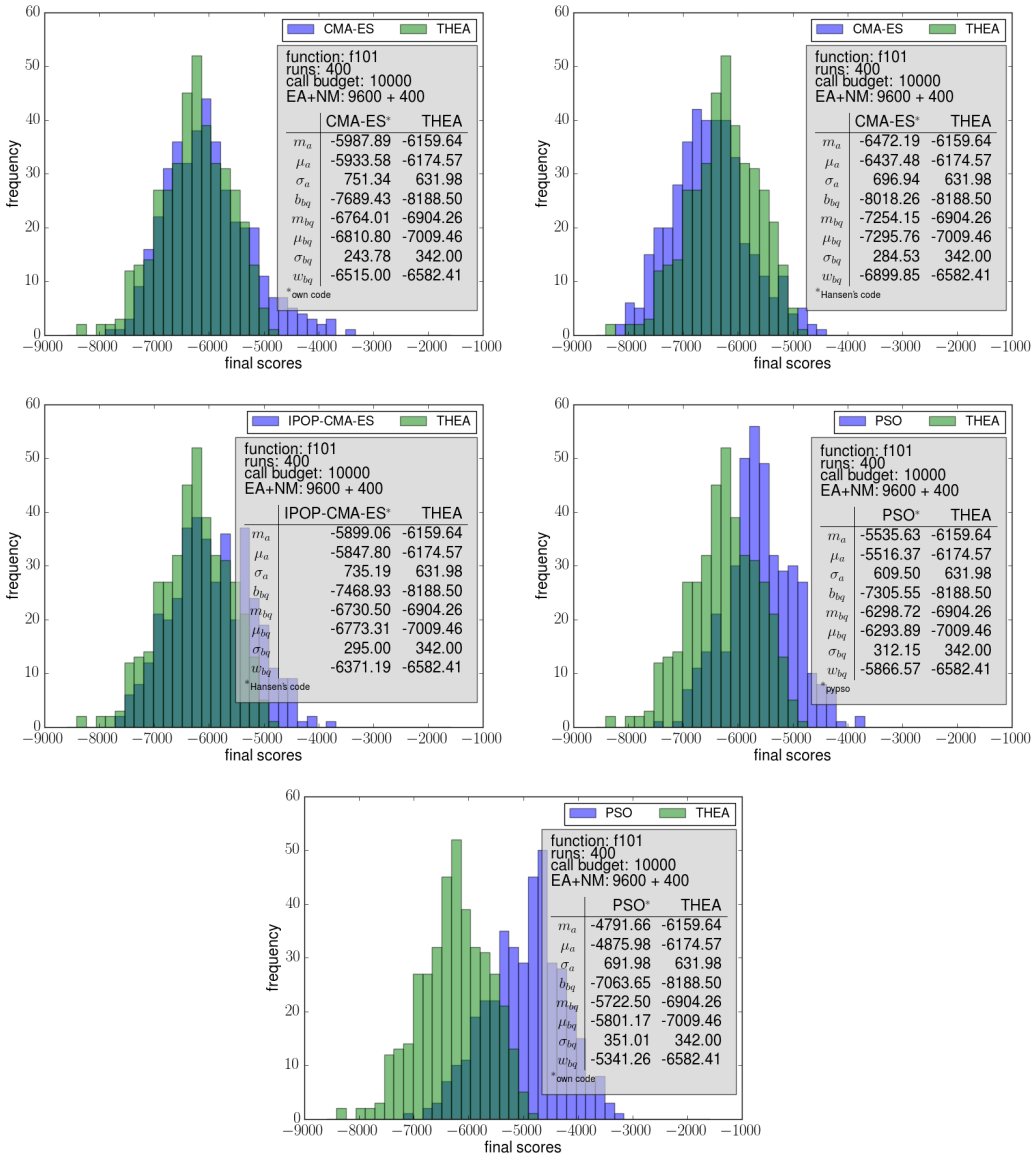


Figure 3.15 Benchmarking THEA on F101 by Whitley et al.
On this test function, only the setup B of CMA-ES is able to beat THEA. PSO handles this search task much worse than the other two EA types.

3.4.4 Interpreting the histograms part 1: the CEC-2005 functions

On the Weierstrass function CMA-ES is simply unbeatable. In each of the tried setups it finds the global minimum so often that the corresponding histogram bars stick out from the otherwise bell-shaped distributions. Of the other EAs, none is able to reach the global minimum except one PSO version very few times. A different picture can be seen with the eF8F2 function. Here, the results are very similar between the three EA types, and few or no hints can be deduced for discriminating between them. The locations and slopes of the left halves of the distributions seem to slightly favour THEA. This is reflected also in the corresponding row of table 3.2 with the Wilcoxon test results. However, CMA-ES, in two of the three setups, is the only algorithm able to find the global minimum. It is just very few hits, but the other EAs have none. The Weierstrass and the eF8F2 functions are hard test functions, but they are not the most difficult challenges to be found in the CEC-2005 collection. However, they have two benefits, for which they were chosen to be part of the benchmark: they produce continuous score distributions (not broken up in levels or bands like in figure 3.14 which is the case for several of the hardest CEC-2005 problems), and they create search landscapes of which it is possible to get an imagination. (What is the meaning of a "good performance" on a test problem of which it is impossible to tell what the similarities and differences are with respect to the real-world problem of interest?) Figure 3.16 shows pictures of the 2D versions of the two functions at different zoom levels. The plots make it clear that the test functions have underlying guiding structures. Larger-scale gradients are hidden underneath smaller-scale ripples, there are structure symmetries with the global optimum at the centre, and in the case of eF8F2 the local valleys get wider and wider towards the deepest one. Since one cannot hope for the conjunction of such beneficial features in the resonator optimisation problem, the benchmark results on these two functions are taken into account with limited weight when making the final EA decision.

However, there is still something to learn from the extremely good performance of the CMA-ES on the self-similar Weierstrass function. The strength that helps CMA-ES here, is its continuous probing the envelope on each level (see figures T.17 & T.16) while sinking down from larger to smaller length scales, and the well-balanced interplay between the tendency to follow the perceived gradient and counteracting forces resembling damping and inertia. The Weierstrass function is a sum of twenty cosines per dimension with exponentially increasing spatial frequencies and exponentially decreasing amplitudes. CMA-ES systematically works off these contributions, beginning with the coarsest and ending with the finest one.

3.4.5 Interpreting the histograms part 2: the three harder problems

Looking at the histograms in figures 3.13 to 3.15, one can observe that in each case one EA sticks out due to either under- or overperformance. In the case of the charged marble problem it is CMA-ES by underperformance. The in-house PSO and THEA perform equally well on the task. Positions and slopes of the left flanks match very closely. In the case of the FM-synthesis it is CMA-ES sticking out but

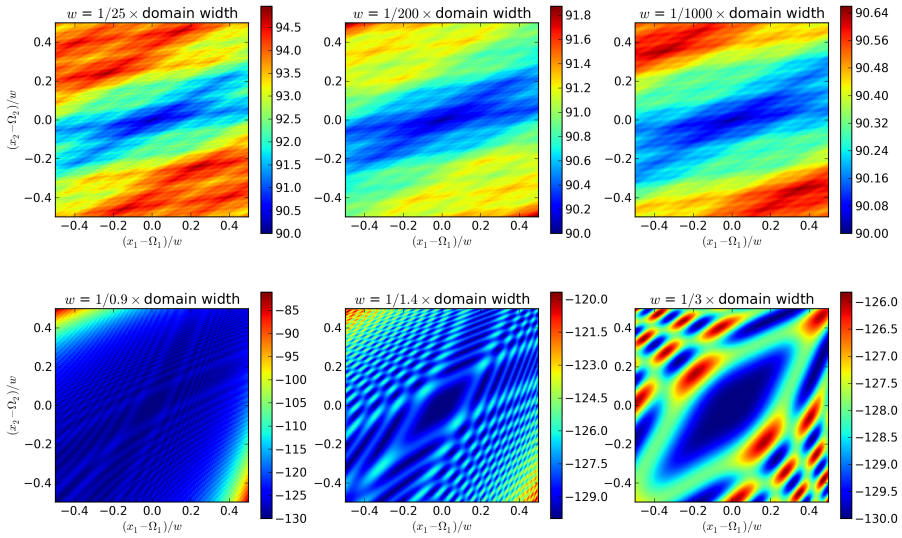


Figure 3.16 Zoom-diving into the Weierstrass and the eF8F2 functions
The CEC-2005 test function suite for EA benchmarking [444] contains many multimodal test functions of varying difficulty level and characteristics. The easier ones are the Rastrigin and Griewank functions (F_9 and F_7). The hardest ones are the hybrid composite test functions made of sums of up to ten shifted and rotated (includes stretching) mathematical test functions. The main problem with these composite test functions is that a human cannot gain any imagination of what type of features make up the largest part of the challenge in the n-dimensional search space. A further problem is that they create broken-up score distributions. Two multimodal functions from the collection, of which it is still possible to gain some imagination, are F_{11}, the shifted rotated Weierstrass function, and F_{13}, the expanded extended Rosenbrock plus Griewank function (historic abbreviation: eF8F2). The formulae are given in appendix V. The above plots show snapshots of zoom dives into the two-dimensional versions of the functions heading towards the position of the global minimum $\vec{\Omega}$. The x- and y-axes cover windows of width w, indicated as fraction of the width of the whole search domain, centred on $\vec{\Omega}$. The top row represents the Weierstrass function and the lower row eF8F2. The most important characteristics that can be inferred for the Weierstrass function are, firstly, that under a superficial ruggedness there is a clear underlying guiding structure leading to the centre of the symmetric valley structure, and secondly, that the landscape is self-similar, i. e. the valley shape looks ever the same and the ruggedness pattern repeats itself while the searcher sinks down through the length scales steering towards the global minimum. The CMA-ES is the ideal algorithm for tackling this type of search challenge because the population can be understood as a probe for the average gradient. Now to the function eF8F2 in the lower row. It exhibits a stretched chess board pattern of hills and valleys. On the global level there is a guiding structure, as the envelopes of both hill tops and valley bottoms are ascending far away from $\vec{\Omega}$. But in the vicinity of $\vec{\Omega}$ this guiding structure is lost because the differences in the levels of summits and bottoms become very small. They would still be detectable and interpretable for a dedicated algorithm characterising hills and valleys locally, but for a search engine relying on samples at random locations, the randomly detected levels will create misleading information if interpreted as average gradient. The problem increases with dimensionality. Only the feature that the valleys become wider towards $\vec{\Omega}$ may still be exploited as a guiding structure because wider valley grounds translate into higher probabilities to sample low function values. (In ad hoc trials, scatter search [170, 172, 249, 287] (code [430], based on [62, 287]) revealed to be quite competitive on eF8F2. Indeed, it seems logic that starting a new local search run somewhere on a connecting line between two valley centres is a beneficial operation for proceeding in steps towards the most central valley of this function's regular pattern of valleys.)

this time due to overperformance again. The histograms are too much broken up due to the separation of bands on the quality axis corresponding to the vicinity of discrete near-optimal solutions. They cannot be compared very well by their shape. The ES is declared the winner here because it finds the global optimum in more than 25 % of the runs, whereas the other two EAs hit it only at a rate of about 15 %. Finally, in the case of F101, it is PSO being the outlier due to underperformance, as it leads to distributions shifted quite a substantial bit to the right.

Daring to judge from the performance on these three deliberately selected hard test problems, THEA can be declared to be a very safe bet because of the fact that it is never the underperforming outlier. This speaks for the robustness of THEA. At the same time it means that the hybrid EA approach has a real pay-off. The benchmark experiments are taken as strong justifying arguments with respect to applying THEA to the resonator optimisation problem. CMA-ES surely is a very good choice, too. It should be tried out by anyone seeking an EA optimiser for an engineering task, unless there is reason for the assumption that the real-world problem has search landscape features with deceiving effect for CMA-ES. PSO seems to be the riskier choice because F101 shows how it can fall back quite substantially behind the other two EAs on a particular problem.[13] Consequently, both CMA-ES and THEA have been applied to the SF resonator optimisation problem.

Comparisons among EA variants

Concerning the comparisons between the variants of CMA-ES and PSO, the statistics still hold some more information, at least for CMA-ES. In the case of PSO, the message is ambivalent because pyPSO is better on the FM-synthesis problem and F101, whereas the in-house PSO runs ahead on the marble problem. About CMA-ES, one important observation comes from the comparison between the barebone algorithm version and the production code equipped with Hansen's latest functionalities and tweaks. While the barebone version finds the global minimum of the 10D Weierstrass function at an incredible rate, the three hard problems reveal that the effort put into the latest[14] CMA-ES version really made the search more global. That there is a price to pay in the form of decreased performance on easier functions like the Weierstrass function when increasing the general-purpose applicability of the searcher, would be in agreement with the "no free lunch" theorem. Additionally, there is something to learn about restarts.

What the histograms tell about restarts

In the case of CMA-ES an interesting comparison is offered between versions with and without restarts of one and the same algorithm. The setup with smaller populations and restarts performs consistently worse than the other two setups based on singular continuous runs with large populations. This may seem surprising considering the background that it was exactly *IPOP-CMA-ES* with its restart strategy

[13] However, it has to be noted that this author has spent the least amount of time with setting up, trying out, and tuning the PSO algorithms. Both examined PSO codes represent standard implementations and thus the literature status of 1995.

[14] the "production code" downloaded from the website [191] in October 2014

that outperformed all competitors in the CEC-2005 competition on real parameter optimisation according the compilation of results [193], and that in 2013 again an EA with CMA-ES and restarts [266] ended up among the top places. The seemingly contradicting facts can be understood by examining the drawback of IPOP-CMA-ES, the composition of the CEC test function suite, and how the EAs have been compared by the CEC jury. IPOP-CMA-ES begins with small populations, and restarts, connected with a doubling of the population size, are initiated upon detecting stalls. For unimodal and easy multimodal functions this means IPOP-CMA-ES can get very far in its low-cost initial setup. However, for challenging multimodal search problems requiring the scanning power of large populations this means large parts of the budget are wasted with less efficient searches and for the last stage with the largest population only a fraction of the total budget remains. In [193] algorithms were ranked higher if they used less of the function call budgets to reach given solution quality targets, and the rankings were based on aggregate data from many test functions. In such a setup it is possible to compensate for decreased performance on the few hardest problems by increased performance on the many easier problems. The statistics here, which represent the question about what has been gained at the end of the budget, show that on the three hard test problems CMA-ES can leverage its global search power much more efficiently if it does not follow the "IPOP" scheme, if the search is conducted with a rather large population size in one single uninterrupted run. This highlights the eternal fact that different experiments answer different questions, and that experiments have to be purposefully designed in order to reveal the desired knowledge. In the context of EAs and engineering problems it means that own EA benchmarking experiments are useful and should be recommended if the benchmarks available in the literature are not framed in the proper setup.

3.4.6 Quantifying the statistics

The decision about the most suitable EA for the SF optimisation task can be settled through the inspection of the score distributions on pages 87 to 89. Nevertheless, some characteristic indicators of quantitative analyses are added to make the presented statistical datasets more useful for readers with a general interest in EA.

Most commonly, the mean, the standard deviation, and the median are used for describing distributions. And advantage of the median over the mean is its insensitivity to outliers. The median separates the two halves of a histogram. Due to the fact that increasing the number of samples taken from a random distribution reduces the likeliness that they are all lying in the same half, the medians of the EA score distributions can be taken as a very conservative measure allowing algorithm decisions based on a pessimistic standpoint saying "you should almost for sure get better results than this if you can afford more than just one shot." In the real-world scenario of tackling a costly optimisation problem there will not be endless repetitions of EA optimisation runs, but also the number of trials will not be as low as just one or two because first attempts are rarely flawless, because the optimisation algorithm, its target simulation, and the interfacing must be set up in steps, and because engineers learn and progress practices in iterations. As it is possible to posit

that we are really only interested in the left half of the distributions, does it make sense to give the median of that part only? Indeed it is deemed worthwhile to be listed because it can be taken as a measure telling "with several shots you're very likely to get at least to this limit." The argument is that the best quarter is still a substantial chunk of the statistics. The chances of missing it n times in a row can be used to calculate the chances of getting something from within the best quarter. The latter probability is

$$P = 1 - \left(\frac{3}{4}\right)^n ,$$

and that is 58 %, 68 %, and 76 % for 3, 4, and 5 trials. With $n = 10$ it's 92 %. Therefore, the best quarter of the statistical raw data gets special consideration in the presented listing of quantitative indicators. Table 3.1 contains this compilation of key numbers distilled from the datasets.

When distributions are not symmetric, not bell-shaped, not continuous, or cumbersome in other ways, then it may still be possible to make meaningful judgements by using *nonparametric statistical tests*. Several such procedures are common in evaluating the performance of optimisation algorithms [114]. The *Wilcoxon signed ranks test*[15] has been applied to check for significant differences in the distribution pairs plotted in figures 3.11 to 3.15. When taking pairs of samples from two sources of random numbers, one can check whether the distribution of pair differences is centred around zero, or whether it is lopsided which would be a hint that behind the two sources there is in fact not the same distribution. The trick is to order the list of incoming pair differences d_i according to absolute size, assign rank 1 to the smallest and rank N to the largest $|d_i|$, and then divide the list of ranks into two piles depending on the sign of d_i. That means rank r_i goes to the winner of that sample pair comparison. The larger one of these two rank sums hints to the winning sample source. By the ranking trick it is avoided that the differences themselves enter the calculation and issues of scaling with them. The smaller rank sum is used to compute a p-value as indicator of the level of significance (some not so trivial ingredients enter that computation [35, 114]). The p-value is the probability that the null hypothesis, that the two datasets stem from the same distribution, is true. Without it, the test is useless. Winning by the rank sums becomes meaningful only in connection with a small p.

The results of the algorithm comparisons by the Wilcoxon signed ranks test are presented in table 3.2. Neither across the whole table nor in any column, row, or other reasonable subsection is there a majority of defeats over wins. This speaks again for the robustness of the presented hybrid EA and for the success of the hybridisation approach.

The weakness of the presented benchmarking tests is that comparisons between THEA and its ingredients like GAs and DE are not included. The problem lies in the vast amount of existing variants of these basic EA schemes. This makes it difficult to choose the strongest competition and too easy to do the other EA ideas a disservice by uninformed choices from among the vast literature and published code versions.

[15] as implemented in the SciPy library [223]

Table 3.1 EA benchmarking statistics

This table compiles analysis data from the statistics in figures 3.11 to 3.15 and allows to benchmark the performance of THEA against CMA-ES and PSO, the former in three different setups, the latter in two. The statistics are compared, on the one hand, according to the median m, mean μ, and standard deviation σ of all runs, denoted by the index a. But, because the fraction of best runs is more relevant with respect to EA decisions in real-world scenarios than the histogram tail of worst runs, the same three quantities are also given just for the best quarter of runs and then indexed with bq. Additionally, the best b_{bq} and worst w_{bq} of the best quarter are listed. While the median m_a represents a very conservative basis for EA decisions (we might get *something* from the better half, but don't count on hitting the better parts of it), the median of the better half (i. e. w_{bq}) allows a still conservative but somewhat less pessimistic decision basis. In particular, going with that criterion means not to ignore the shape of the better halves of the distributions. Comparing the EAs by w_{bq}, one can see that on the three hard problems THEA is not beaten that often: on the charged marble problem none is better, and on both the FM-synthesis problem and F101 only one of the ES variants, CMA-ES-B, gets ahead.

	m_a	μ_a	σ_a	b_{bq}	m_{bq}	μ_{bq}	σ_{bq}	w_{bq}
				charged marbles (8D)				
CMA-ES-A	31.052	31.107	0.545	30.237	30.528	30.507	0.123	30.695
CMA-ES-B	30.988	31.082	0.614	30.165	30.484	30.470	0.100	30.634
CMA-ES-C	30.972	31.068	0.511	30.201	30.542	30.514	0.118	30.702
PSO-A	30.951	31.059	0.597	30.140	30.479	30.472	0.116	30.647
PSO-B	30.727	30.797	0.365	30.230	30.420	30.414	0.071	30.533
THEA	30.695	30.769	0.347	30.183	30.414	30.399	0.078	30.497
				FM-synthesis (6D)				
CMA-ES-A	11.512	10.293	6.239	0.000	0.000	1.282	1.700	5.364
CMA-ES-B	8.444	9.310	7.385	0.000	0.000	0.000	0.000	0.002
CMA-ES-C	13.725	12.137	6.578	0.000	3.328	2.976	2.775	8.440
PSO-A	11.512	10.040	6.199	0.000	0.000	1.053	1.577	5.364
PSO-B	15.486	14.355	5.086	0.000	9.334	7.251	4.215	11.512
THEA	12.152	10.251	6.142	0.000	0.000	1.501	2.139	5.364
				F101 (10D)				
CMA-ES-A	-5987.9	-5933.6	751.3	-7689.4	-6764.0	-6810.8	243.8	-6515.0
CMA-ES-B	-6472.2	-6437.5	696.9	-8018.3	-7254.1	-7295.8	284.5	-6899.9
CMA-ES-C	-5899.1	-5847.8	735.2	-7468.9	-6730.5	-6773.3	295.0	-6371.2
PSO-A	-5535.6	-5516.4	609.5	-7305.6	-6298.7	-6293.9	312.2	-5866.6
PSO-B	-4791.7	-4876.0	692.0	-7063.7	-5722.5	-5801.2	351.0	-5341.3
THEA	-6159.6	-6174.6	632.0	-8188.5	-6904.3	-7009.5	342.0	-6582.4
				Weierstrass (10D, only error given)				
CMA-ES-A	0.004	0.084	0.327	0.000	0.001	0.001	0.000	0.002
CMA-ES-B	2.128	3.006	2.908	0.004	0.077	0.227	0.263	0.902
CMA-ES-C	2.210	2.638	2.120	0.011	0.509	0.533	0.437	1.201
PSO-A	4.993	4.910	1.570	0.948	3.023	2.880	0.616	3.646
PSO-B	3.882	4.592	2.789	0.192	1.517	1.478	0.569	2.330
THEA	2.580	2.628	1.046	0.428	1.284	1.330	0.359	1.905
				eF8F2 (10D, only error given)				
CMA-ES-A	0.973	0.974	0.267	0.474	0.703	0.677	0.073	0.759
CMA-ES-B	0.882	0.893	0.261	0.010	0.640	0.587	0.145	0.715
CMA-ES-C	0.797	0.836	0.253	0.040	0.564	0.544	0.085	0.649
PSO-A	0.867	0.919	0.357	0.271	0.551	0.533	0.103	0.666
PSO-B	0.870	0.916	0.325	0.273	0.566	0.552	0.085	0.671
THEA	0.782	0.824	0.274	0.235	0.543	0.515	0.094	0.636

Table 3.2 Wilcoxon test on EA statistics

The Wilcoxon signed ranks test [114, 223] has been applied to the pairs of solution quality distributions in figures 3.11-3.15. This nonparametric test for comparing two sequences of samples is based on pairwise comparisons. The question is whether the distribution of pair differences is symmetric around zero. The listed p-values indicate the probability of the null hypothesis to be valid, i. e. that the samples come from the same distribution, that the difference distribution is by consequence symmetric, and that the measured asymmetry stems only from randomness in the sampling. The comparisons have been declared a tie in the cases of $p > 0.1$. There is one case where p is very close to 0.1, and this case would be a defeat if counted.

	CMA-ES-A	CMA-ES-B	CMA-ES-C	PSO-A	PSO-B
			Weierstrass (10D)		
THEA wins?	defeat	tie	tie	win	win
p-value	5.05×10^{-67}	6.40×10^{-1}	1.03×10^{-1}	8.48×10^{-57}	4.13×10^{-26}
			eF8F2 (10D)		
THEA wins?	win	win	tie	win	win
p-value	4.83×10^{-13}	4.07×10^{-5}	4.33×10^{-1}	8.28×10^{-5}	4.39×10^{-5}
			charged marbles (8D)		
THEA wins?	win	win	win	win	tie
p-value	7.33×10^{-21}	1.66×10^{-14}	5.01×10^{-17}	1.41×10^{-14}	2.39×10^{-1}
			FM-synthesis (6D)		
THEA wins?	tie	defeat	win	tie	win
p-value	9.72×10^{-1}	5.02×10^{-2}	7.80×10^{-5}	5.06×10^{-1}	4.69×10^{-21}
			F101 (10D)		
THEA wins?	win	defeat	win	win	win
p-value	1.80×10^{-5}	4.79×10^{-9}	1.93×10^{-9}	4.70×10^{-33}	4.39×10^{-62}

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
a_i	lower bound of search space along i^{th} dimension
b	branching factor of population merging scheme
b_a, b_{bq}	best of all samples = of best quarter (benchmark statistics)
b_i	upper bound of search space along i^{th} dimension
c_1, c_2	attractor force scaling factors in PSO description
f_{obj}	objective function
f_{sel}	parent selection function implementing selection pressure
G	maximum number of generations (algorithm control)
g	generation counter
m_a, m_{bq}	median of all samples, of best quarter (benchmark statistics)
N	population size
n	search space dimension
P	probability
$\mathcal{P}$	chromosome population
p	selection pressure control parameter
$\mathbb{R}$	real numbers
r	random number
s	scaling factor for DE routine
w	window width of test function zoom plots

w_a, w_{bq}	worst of all samples, of best quarter (benchmark statistics)
$\vec{x}$	chromosome, i. e. point in search space
$\{\vec{x}\}$	set of chromosomes

List of Greek quantity symbols

Symbol	Description
α	inertia parameter in PSO description
α	recombination operator extension range beyond better parent
β	recombination operator extension range beyond worse parent
γ	annealing factor (mutation step size reduction)
θ	generic transformation function for distribution forming
ϑ_{c2u}	"cigar-to-uniform" ratio of recombination operator usage
κ	mutation step size scaling factor
λ	size of offspring population
μ	size of parent population (selected for "survival" and reproduction)
μ	mean value
μ_a, μ_{bq}	mean of all samples, of best quarter (benchmark statistics)
ν	integer output variable of parent selection routine
σ	mutation step size control parameter
σ_a, σ_{bq}	std. dev. of all samples, of best quarter (benchmark statistics)
ψ	attractor force scaling factor in PSO description
$\vec{\Omega}$	location of global optimum

List of abbreviations

Abbreviation	Description
CEC	IEEE Congress on Evolutionary Computation
CMA-ES	evolution strategy with covariance matrix adaptation
CO	crossing-over, crossover
DE	differential evolution
DNA	deoxyribonucleic acid
EA	evolutionary algorithm
EAO	evolutionary algorithm optimisation (meaning optimisation by evolutionary algorithm)
EC	evolutionary computation
ES	evolution strategy
FE,FEM	finite element (method)
FM	frequency modulation
GA	genetic algorithm
ILS	iterated local search
IPOP-CMA-ES	restart CMA-ES with increasing population size
LS	local search
NM	Nelder-Mead algorithm, i. e. downhill simplex search
KIT	Karlsruhe Institute of Technology (Karlsruher Institut für Technologie)

PSO	particle swarm optimisation
RLS	repeated local search
RPI	Rensselaer Polytechnic Institute
SA	simulated annealing
SF	sonofusion
THEA	tier-based hybrid evolutionary algorithm
WHX	Wright's heuristic crossover

Chapter 4

Interfacing optimiser and simulation

This chapter discusses the aspect of finding a suitable objective function giving the global EA search the right direction so that parameter combinations yielding a maximised fitness computation do indeed represent good solutions to the engineering problem. The multi-step simulation and postprocessing routine for evaluating a resonator solution candidate is motivated and explained. Different setups of the fitness evaluation scheme are presented, which are at the basis of the different EAO results presented in the following chapter, and which are able to determine whether the global evolutionary search is applied either in a more explorative or in a highly targeted manner.

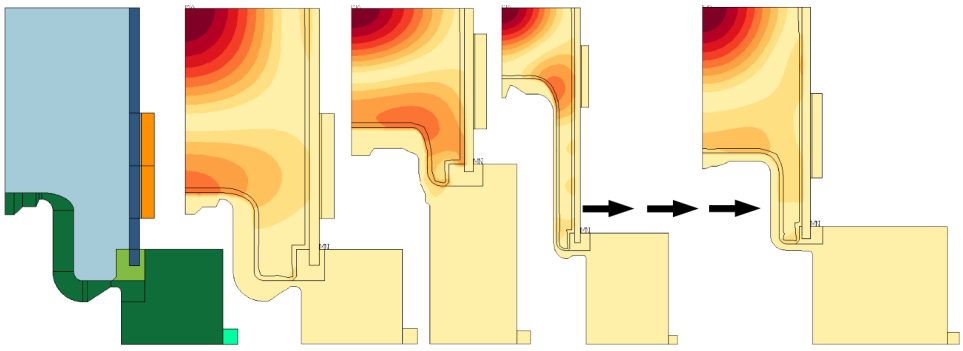


Figure 4.1 Turning random trials into good SF resonators
This collection of resonator model cross sections shows three random trials on the left and a highly optimised setup on the right. The colour contours represent normalised pressure amplitudes. The random-influenced geometry setups are from an early stage of EA optimisation and they represent solutions of low quality because they exhibit either a low peak sound pressure or a high pressure amplitude on walls or both in conjunction. The values for the four setups are 24.1, 8.6, 29.5, and 39.6 bar for the central pressure and the maximum pressures on walls are at 53, 49, 51, and 24 % of the central pressure. A well-performing global search algorithm is necessary for achieving this optimisation result, but the other crucial part is a suitable objective function definition guiding the search towards good solutions of the engineering problem.

4.1 General considerations about the optimisation goals

4.1.1 The question of multi-objective or single-objective search

The goals of the SF resonator design problem were formulated in chapter 2. Using an automated procedure of parameter tuning for design optimisation means that the list of goals has to be split into two halves: the goals served when making the design concept and those goals that are left up to the automatic optimiser. Clearly, the topics of reproducibility, machining, assembly technique, material choices etc. manifest themselves in the first half. A comprehensive list of goals offering themselves as part of the second half could look like this:

1. maximise the pressure amplitude in the centre,
2. minimise the pressure along the fluid-structure boundary,
3. maximise the Q-factor,
4. minimise material stresses,
5. maximise the displacement amplification mechanism if it is part of the vibration mode,
6. minimise the displacement near nozzles and tube connections,
7. minimise the deformation where there are silicone connections in order to avoid excessive damping,
8. maximise the volume size suitable for spherical bubble cloud collapse, i. e. prefer pressure mode shapes with fewer nodes.

Of course, it is not possible to treat the eight goals separately and equally per multi-objective (MO) search because of the curse of dimensionality, the difficulty to make any use of the ensemble of solutions forming the 7D Pareto front in the 8D space, and the exploding number of trials needed. But still, the MO versus SO (single-objective) question can be asked more seriously in this form: should the optimisation perhaps better be made as a bi- or tri-objective search concentrating on e. g. the first, second and/or fourth goal in the above list? The decision against this option was partially made for avoiding the increased computational cost connected with more necessary trials but also because MO optimisers, their inner workings, their emergent properties, and their benchmarking are yet one more nontrivial topic.

Proceeding with SO optimisation requires the condensation of the few most important and most independent goals into one single objective function of scalar value. Some might be ignored because of interrelations, e. g. a low Q-factor will not allow high pressure amplitudes. Others have to be addressed as much as possible through suitable design (pushing into the first half) or by bounding the parameter space (goal represented in second half but not through f_{obj}), if they cannot enter the objective function formulation. In all SF resonator optimisation runs discussed below, single-objective optimisation has been applied with a scalar objective function based on solely the first two goals in the above list. If multiple goals are condensed into

one scalar objective function, then it is of great importance how exactly they get merged[1] together because the search landscape determines how successful a search algorithm can be.

4.1.2 Taking care on the fitness function avoids pitfalls

When EAs or other function optimisers are applied to real-world problems, in particular from the field of engineering, then the formulation of the objective function is a crucial point. The objective or fitness function connects the phenotype solution with the algorithm, and if no care is taken for the setup of this connection, then two types of frustrations are likely to occur:

1. solutions are found which maximise the fitness indeed but represent no satisfying solution to the engineering problem because of another criterion, or
2. the search efficiency is low, many mediocre and too few good solutions are found because the topology of f_{obj} hides the really good solutions from the searcher.

Imagining for the moment the simplest case, that the scalar objective function to be maximised is just the peak sound pressure anywhere near the central region of the resonator (the resonance being anywhere in the scanned frequency interval), then a problem of the first type might show up in the form that the desired pressure increase in the middle is accompanied by high pressures somewhere at a fluid-structure boundary or connected with excessive stress concentrations in the structure. A problem of the second type could reveal itself in a series of optimisation runs where in only one of several cases a distinctly different solution of much better quality than in all the other cases is found. In the resonator case where solutions can be classified by their pressure mode shapes it is thinkable that solutions with certain types of mode shapes usually take over the gene pool in the early stages of evolution, so that the one type of mode shape allowing much higher pressures just never gets efficiently searched and locally optimised. The situation could also be described in the following way: the attractor regions of the different working modes are of very unequal size in the search space. In such a situation, if the user is able to decide ad hoc which solution type is preferable, penalising the fitness of undesired solution types can be an efficient way for substantially shrinking the attractor regions of the solutions declared wrong.

In fact, the dissection between the two problem types is not that clear. This can be made clear by interpreting the examples for the first type also with the help of the

[1] There is an option of serving multiple goals without merging them into a single scalar objective function: by swapping the goals every other generation and staying with an SO evaluation in each generation (see appendix T.5). For example, the default pressure-based goal of the resonator optimisation could be exchanged in every n^{th} generation by a stress-based goal. Experiments have been made with the charged marble problem where the two types of energies can be treated as two independent goals. These experiments remained however inconclusive because there were not enough of them and because more work would have been necessary for elaborating sound definitions of measures to evaluate the pseudo-MO algorithm and compare it to the SO alternative, performance measures that would allow to make a decision based on the test problem and carry it over to the resonator optimisation.

attractor image: perhaps there is a design with top central sound pressure and at the same time low interface sound pressure but it is just never found because there are many, many more solutions with high interface pressure representing the haystack burying the needle. It might be useless and fruitless to ponder which problem type might exist to what degree, but it is surely helpful to turn the two thoughts into two tasks that have to be fulfilled by the objective function:

1. The objective function should efficiently block all kinds of undesired solutions. If there are other judgement criteria than just the peak sound pressure, then they should be reflected in the scalar fitness values. If this is to be done efficiently, then the search algorithm should be prevented from being able to over-compensate bad performance on one criterion with extremely good performance on another one.

2. Besides efficiently discriminating between desired and undesired solutions, the objective function has the task to shape smooth transitions in the search space. It should also indicate the neighbourhoods of good solutions. One should ask whether measures can be taken which enlarge the attractor regions of desired solutions and shrink other attractors.

These thoughts are connected to an argument made in the EA background appendix chapter with figure T.3 (p. 481). A practical consequence is for example to avoid a fitness function of the following form: the fitness value of a candidate solution is by default $p_{\max}$, but it is overwritten with the number -1 if one of several checks on other criteria fails. Such a form would rob large portions of the search landscape of any guiding structure. Subtracting penalties from $p_{\max}$ is better. Smooth penalty functions are best.

4.2 The implementation

As part of this work, only pressure amplitudes have been used for deriving the fitness functions. The approach taken to address the other goals was to learn from the solutions found by EAO and to incorporate the new knowledge into the next design step and optimisation run. That means, to react to other undesired developments, e. g. excessive local deformation and stress concentration, by manually setting proper boundaries on the design parameters or introducing fillets for preventing these features. Therefore, the fitness of solution candidates is only based on sound pressure data gathered from the FEM simulations.

4.2.1 A two step approach of resonance identification and characterisation

Each FEM simulation of a resonator is a forced harmonic analysis covering a frequency range in a number of discrete steps. The computations are based on parametrised APDL scripts. In order to be efficient, the evaluation of one FE model has to be based at least on two forced-harmonic frequency sweeps. The first sweep is for covering the whole frequency band of interest in a coarse discretisation. It just

has to be fine enough to be able to choose the most promising resonance most of the times. The second sweep has to focus on the chosen best resonance in a much finer discretisation along the frequency axis, so that the resonance peak is well-resolved and that there is no discretisation error on the pressure peak computation. (In some cases the routine was based on three frequency sweeps allowing for a shifted repetition of the wide scan in cases where the peak was located on the borders.) The used APDL scripts contain routines for looping over selected nodes and writing nodal solution data into text files (employing the command `*vwrite`). Based on that data, the Python-based postprocessing and score evaluation routines can analyse the data step by step and compute the fitness as illustrated in figures 4.2, 4.5, and 4.5.

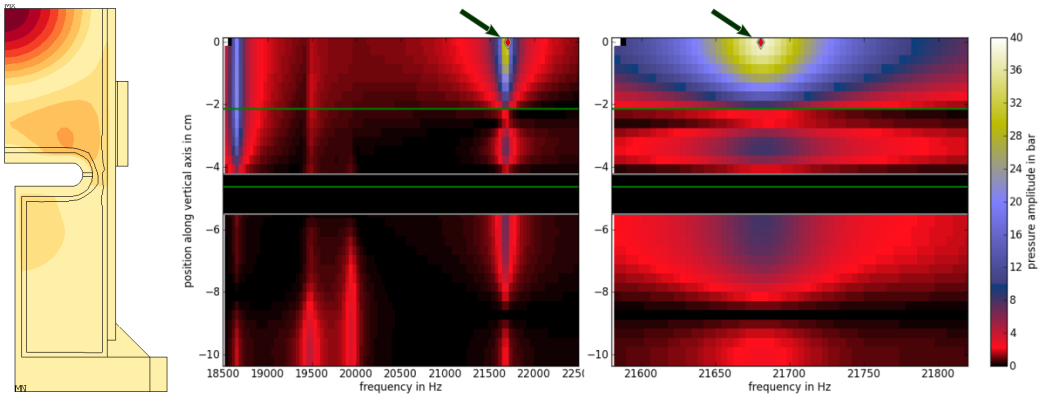


Figure 4.2 Fitness determination in two steps
These plots show pressure amplitude maps gained from FEM simulations. In the left picture pressure amplitude colour contours are shown across the FE mesh of the modelled lower half of a symmetric resonator variant, depicted is the situation at resonance. The colour maps to the right show the pressure amplitude profiles along the central axis of the resonator over the frequency axis. They are gained with Python-based postprocessing routines analysing the FEM output. The z-axis range covered by the piston is blacked out in the map. The plotted data behind the piston is the pressure amplitude measured on the surface of the piston holding tube. The two frequency sweeps, one wide scan and one zoomed-in close-up of the automatically detected strongest resonance at 21.7 kHz represent the two step solution candidate evaluation procedure. In the simplest (earliest developed) evaluation scheme it is only the peak pressure in the resonator centre p_{centre} which counts for the fitness.

The earliest trials of EA optimisation of resonator models were attempted with a fitness function aiming solely of the maximisation of the peak sound pressure. An exemplary optimisation result is shown in figure 4.2 it shows the final resonator geometry setup and how its evaluation looks like. A first frequency sweep covers a wide range, here it's a window of 3 kHz. The most promising resonance is identified by searching for the largest pressure amplitude value in the central part of the resonator in between the two pistons on the central axis (denoted in the following as p_{centre}). As profiles with pressure nodes near the pistons are desired, it is beneficial to restrict this scan to a window along the vertical axis excluding the space near the pistons. The red diamond (and the added arrow) indicates the resonance chosen by the evaluation routine. The map to the right is the 240 Hz wide close-up taken in 60 steps around this resonance for characterising it. The data resolution in the close-up map has to be sufficiently fine along the spatial as well as the frequency axis in order to be able to infer the height of pressure amplitude peak with decent accuracy. In the depicted case the frequency step is 4 Hz and the vertical probing

grid takes samples from every second mesh node; with a mesh resolution setting of 1.5 mm this yields a grid resolution of 3 mm.

The simple evaluation routine aiming at the maximisation of p_{centre} without any further constraints (except a restricted scan window along the spatial z coordinate) yielded some interesting optimised resonator setups, but overall it turned out to be unreliable. The reason was that too often the achievement of a high central peak pressure was accompanied by elevated sound pressure amplitudes on fluid-structure interfaces. The problem occurred that chromosomes won the evolutionary competition which corresponded to resonators like the undesired case shown in figure 4.3.

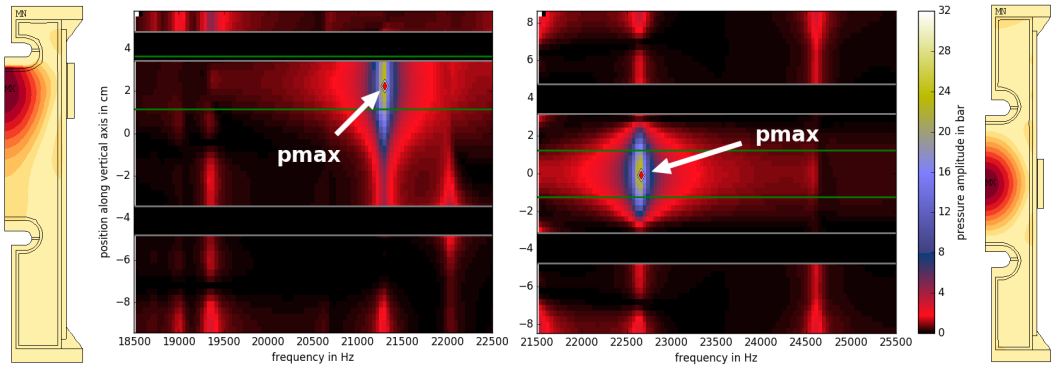


Figure 4.3 The importance of filtering out undesired pressure profiles
These plots illustrate a comparison between bad (left side) and good (right side) pressure profiles. Both resonator setups achieve high peak sound pressures above 20 bar. The problematic aspect of the setup on the left is the elevated sound pressure amplitude right in front of the upper piston. Such a resonator is not suitable for SF trials because cavitation is likely to occur on the surface of the piston and not deep in the liquid volume. Secondly, the condition of spherical symmetry of the sound field is not given for the region where cavitation is expected. This will not allow the concentric collapse of spherical bubble clusters.

If experience implies that it is much easier for the tried EAs to come up with such unsuitable solutions with pressure peaks near or at walls than with suitable ones if the search is simply aimed at increasing p_{centre} then what are the reasons? The behaviour can be made plausible by considering the following thoughts: If the structural parts are only made of very thin walls, then the boundary conditions (BCs) around the fluid domain will in most places be low pressure and high displacement amplitude. Sturdy structural parts will usually entail BCs where it is the other way round. Hard, non-vibrating walls attract sound pressure antinodes. However, unreasonably thin structural parts as the easy way out are intentionally excluded through the setting of parameter bounds. One big advantage of sturdy setups in the evolutionary competition is low damping due to little displacement because damping only occurs in the structure in the considered FE models. Relatively massive structural parts need special conditions to set them in motion, e. g. a structural resonance in conjunction with an acoustic resonance inside the resonator. This is exactly what is needed if the walls are to serve as low pressure amplitude BC while not becoming unrealistically thin. The search for these special conditions is the main part of the optimisation task of the SF resonator design problem. An automatic optimisation routine can be trimmed to accomplish this task by telling it

to exclude areas close to structures on the search for p_{centre} and to penalise elevated pressure amplitudes on fluid-structure interfaces. Implementing such measures in the score evaluation subroutine means shrinking the large and blinding attractor regions of unsuitable designs in the search space and letting the attractor regions of suitable solutions shine through.

4.2.2 Penalising undesired pressure amplitude distributions

The basic idea was to compose the objective function f_{obj} as the sum of p_{centre} and a penalty given depending on the highest pressure amplitude detected anywhere along the fluid-structure interface p_{if},

$$f_{\text{obj}} = p_{\text{centre}} - f_{\text{pen}}(p_{\text{if}}).$$

But first trials with a penalty of the form

$$f_{\text{penA}} = \begin{cases} p_{\text{if}} - p_{\text{lim}} & \text{if} \quad p_{\text{if}} > p_{\text{lim}} = \frac{1}{5} p_{\text{centre}}, \\ 0 & \text{else} \end{cases}$$

did not work because solutions won the competition and conquered the gene pool by overcompensating the penalty with elevated values of p_{centre}. Subsequent trials with a stepwise steeper and steeper slope of the penalty

$$f_{\text{penB}} = \begin{cases} 0 & \text{if} \quad p_{\text{if}} \leq p_{\text{lim},1}, \\ \alpha_1 (p_{\text{if}} - p_{\text{lim},1}) & \text{if} \quad p_{\text{lim},1} < p_{\text{if}} \leq p_{\text{lim},2}, \\ \alpha_1 (p_{\text{if}} - p_{\text{lim},1}) + \alpha_2 (p_{\text{if}} - p_{\text{lim},2}) & \text{if} \quad p_{\text{if}} > p_{lim,2}, \end{cases} \tag{4.1}$$

whereby $p_{\text{lim},1} = \frac{2}{10} p_{\text{centre}}$, $\alpha_1 = 2$, $p_{\text{lim},2} = \frac{3}{10} p_{\text{centre}}$, and $\alpha_2 = 5$, were still problematic because too many promising solutions and their mutations were thrown into the negative fitness range. In this case, the lowered efficiency of the search arose from bad solutions featuring the lowest values of p_{centre} populating the higher ranks of the population above many penalised solutions with originally very promising pressure profiles of strong amplitude.

An approach of quenching the fitness of candidates with far too high wall pressures instead of throwing them into the negative fitness range, and thus behind all other candidates, turned out to be more efficient. A reduction of the fitness down to zero but not further can be achieved if the penalty is made a multiplier, i. e.

$$f_{\text{obj}} = p_{\text{centre}} \cdot f_{\text{pen}} \qquad \text{with} \qquad f_{\text{pen}} : \ \mathbb{R} \to [0, 1].$$

In need of a smooth transition function for bridging the region between strongly penalised and not penalised solutions, one offering a simple tuning parameter for the width of the transition region, the choice fell on the function describing Fermi-Dirac statistics:

$$f_{\text{FD}} = f_{\text{FD}}(x, \mu, \beta) = \frac{1}{\exp\left(\frac{x-\mu}{\beta}\right) + 1}, \qquad f_{\text{FD}}(x) : \ \mathbb{R} \to [0, 1].$$

After defining the wall pressure ratio

$$r_{\mathrm{wp}} = \frac{p_{\mathrm{if}}}{p_{\mathrm{centre}}}$$

the fitness function can be written as

$$f_{\mathrm{obj}} = p_{\mathrm{centre}} \cdot f_{\mathrm{penC}} \tag{4.2}$$

$$= p_{\mathrm{centre}} \cdot (1 - \tilde{f}_{\mathrm{penC}}(r_{\mathrm{wp}})) \tag{4.3}$$

with

$$\tilde{f}_{\mathrm{penC}} = (1 - f_{\mathrm{FD}}(r_{\mathrm{wp}}, \theta, \beta_1))\,(1 - f_{\mathrm{FD}}(r_{\mathrm{wp}}, \theta/3, \beta_2)) \tag{4.4}$$

and

$$\theta = 0.5, \quad \beta_1 = 0.15, \quad \beta_2 = 0.08.$$

Here, the threshold p_{lim} has been replaced by the dimensionless number θ. The three development steps of the penalty function are illustrated in figure 4.4.

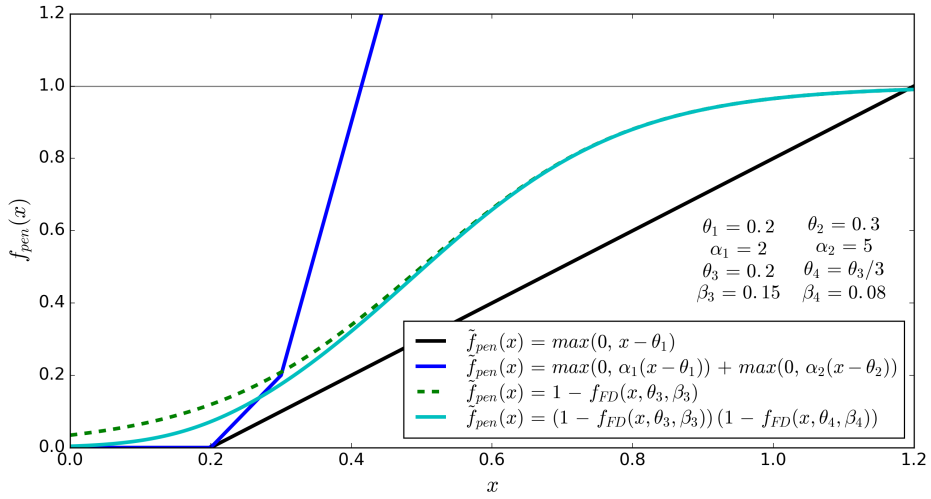


Figure 4.4 Developing the penalty function punishing elevated wall pressure ratios
The three different versions of the penalty function are shown in a form allowing the usage according to $f_{\mathrm{obj}} = p_{\mathrm{centre}}(1 - \tilde{f}_{\mathrm{pen}}(r_{\mathrm{wp}}))$. The problem with the two piecewise linear penalty functions is that either the penalty is not strong enough or it is too strong once it creates the situation of zero or negative scores. The smooth transition function based on f_{FD} offers the advantages that the penalty increases smoothly and steeply in a decisive region which can be shifted by the parameter θ, but at the same time, only in cases of extreme wall pressure ratios the fitness will be decreased below the one of candidate solutions with very low peak pressure amplitudes in the centre. Typically, in the bulk of the population, the peak sound pressure is seen to vary by a factor of five to ten.

4.2.3 A simple score evaluation routine without mode shape discrimination

Figure 4.5 shows two colour maps with raw data, the sound pressure amplitude along the central axis (top) and measured along the inside of the outer hull (second map). Clipping the centreline pressure map at top and bottom and ignoring the spaces close to the pistons already helps a little bit in gearing the optimiser to push

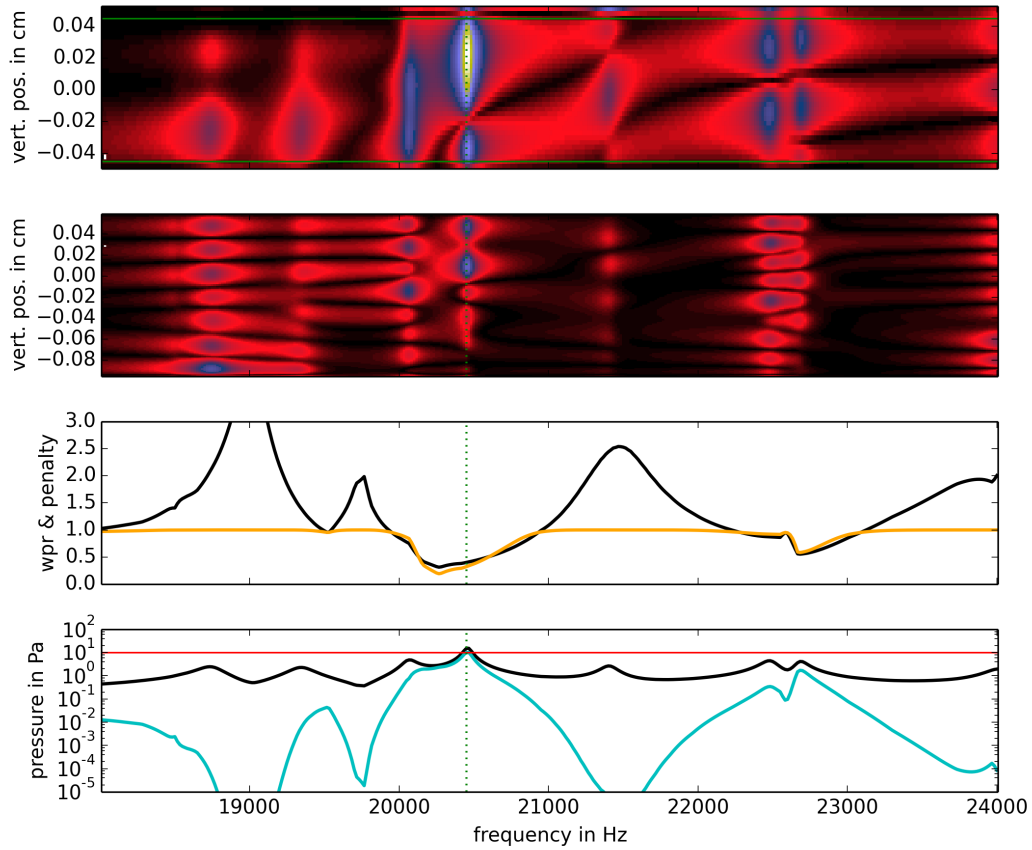


Figure 4.5 Score evaluation routine A: mode-independent version
This combined diagram gives an overview of the whole data postprocessing and score evaluation routine. The two colour maps represent the sound pressure along the central axis over frequency (top) and along the inside of the main glass wall (second from top). A region of interest along the central axis is indicated by green lines, it excludes the space close to piston surfaces. Based on the maximum pressure in the central region of interest p_{centre} and the maximum pressure on fluid structure interfaces p_{if} the wall pressure ratio can be calculated (black curve in the third diagram). The penalty function $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{centre}}, p_{\mathrm{if}})$ is depicted in the same diagram in orange. Lastly, the fourth plot illustrates the score computation showing the peak centreline pressure p_{centre} over the frequency in black and the penalised objective function $f_{\mathrm{obj}} = p_{\mathrm{centre}} \cdot f_{\mathrm{pen}} = p_{\mathrm{max}}(f) \cdot (1 - \tilde{f}_{\mathrm{pen}}(f))$ in cyan. (The depicted exemplary frequency response is from the FEM simulation of the West-Howlett resonator in the setup shown in figure Q.16. It has a pressure peak of 15.2 bar and a wall pressure ratio of 40 %. The penalisation scheme reduces the score from 15.2 to 10.3 given a threshold setting of $\theta = \frac{1}{2}$.)

for mode shapes with antinodes in the inner region. The score is based on the peak centreline pressure p_{centre} from that region of interest and the intention is to penalise it when there are high pressure amplitudes p_{if} anywhere along fluid-structure interfaces through the wall pressure ratio-based penalty function. The computation of p_{if} involves scanning more raw data than depicted in the two pressure maps, in particular the pressure data covering the whole piston surfaces (see figure 4.6 depicting the consequences of an erroneous computation).

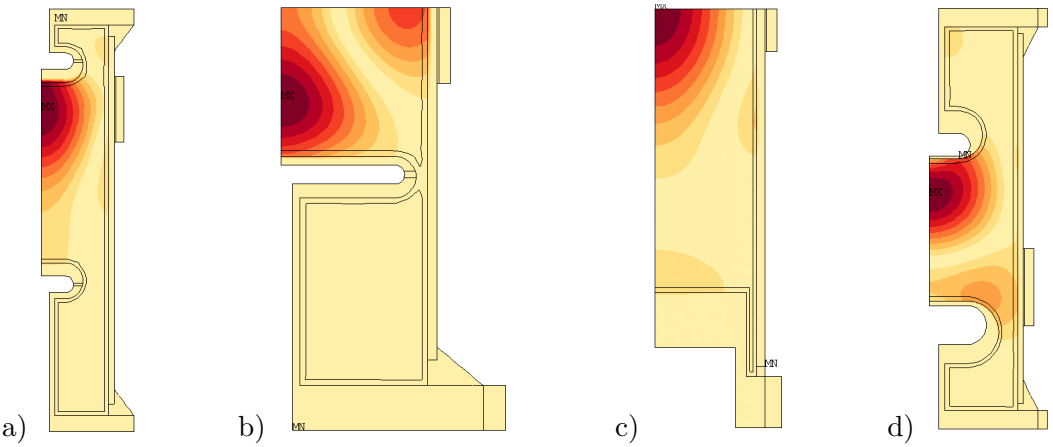


Figure 4.6 The effect of a flawed evaluation routine
The calculation of the wall pressure ratio and the penalty should take into account the entire fluid-structure interface area. Unfortunately, a flawed computation was implemented in several of the optimisation runs presented in chapter 5. Only the top and bottom lines of the centreline pressure map and the pressure map from the main glass wall (as plotted in the top two diagrams of figure 4.5) were taken into account. This means data measured on the middles of the piston front plates were assumed to be representative of the pressure maxima across all piston surfaces, and signals from the remaining interface areas, e. g. base plates, were assumed to be generally smaller in amplitude. These assumptions were based on early optimisation results like mode shapes (a), (b), and (c), depicted above. At least the first assumption was proven wrong when later optimisation runs yielded mode shapes like in setup (d) shown above. Here, it can be seen that the membrane-like thin piston front plates, because they are softer in the middle and enable less displacement near the rim, have the consequence of leading to concave pressure amplitude isocontours in front of the pistons. In such cases the interface pressure amplitude is higher near the piston rim and measurements taken solely in the front plate centres miss the maxima. The effect is visible on both the upper and lower piston in case (d). The wall pressure ratio and penalty computation routine needed to be corrected for later optimisation runs.

The third plot in figure 4.5 shows the wall pressure ratio $r_{\text{wp}} = p_{\text{if}}/p_{\text{centre}}$ in black and the derived penalty function $\tilde{f}_{\text{penC}}$ in orange. The function $\tilde{f}_{\text{penC}}$ is then used to calculate a score-over-frequency curve $f_{\text{obj}}(f)$ according to equation 4.3, that transformation is shown in the fourth plot of figure 4.5. In this case $p_{\text{centre}}(f)$ and $f_{\text{obj}}(f)$ have their global maxima on the same frequency point, but this cannot be relied on. Therefore, calculating the score of a design only at the global peak of $p_{\text{centre}}(f)$ will not be the most effective evaluation routine because it will miss weaker resonances yielding better scores due to lower penalties.

4.2.4 Tighter targeting with mode shape discrimination

As the result gallery in the next chapter will show, the penalised goal function described above leads to finding various local minima of high performance and exhibiting a multitude of different mode shapes. It shows the usefulness of the EA-based

global optimisation approach for exploring the design space, identifying superior local minima quickly, and finding interesting new setups which a targeted manual search might probably miss.

Which types of different suitable mode shapes a resonator design can host was however not the only question during EAO trials. For each one of the three new investigated geometries discussed in chapter 5 one main question was: how well it can perform when it works in the mode shape known from the West-Howlett resonator, the asymmetric mode shape with an out-of-phase pressure peak well above the transducer allowing for mechanical displacement amplification along the main glass wall? Trying to answer this question, it can be tedious if only one in four EAO runs or less yield the thought-of mode shape. This created the desire to implement a modified solution candidate evaluation routine helping to gear the optimiser towards realising a resonator setup which reproduces the mode shape of the West-Howlett design. The developed scheme is illustrated in figure 4.7.

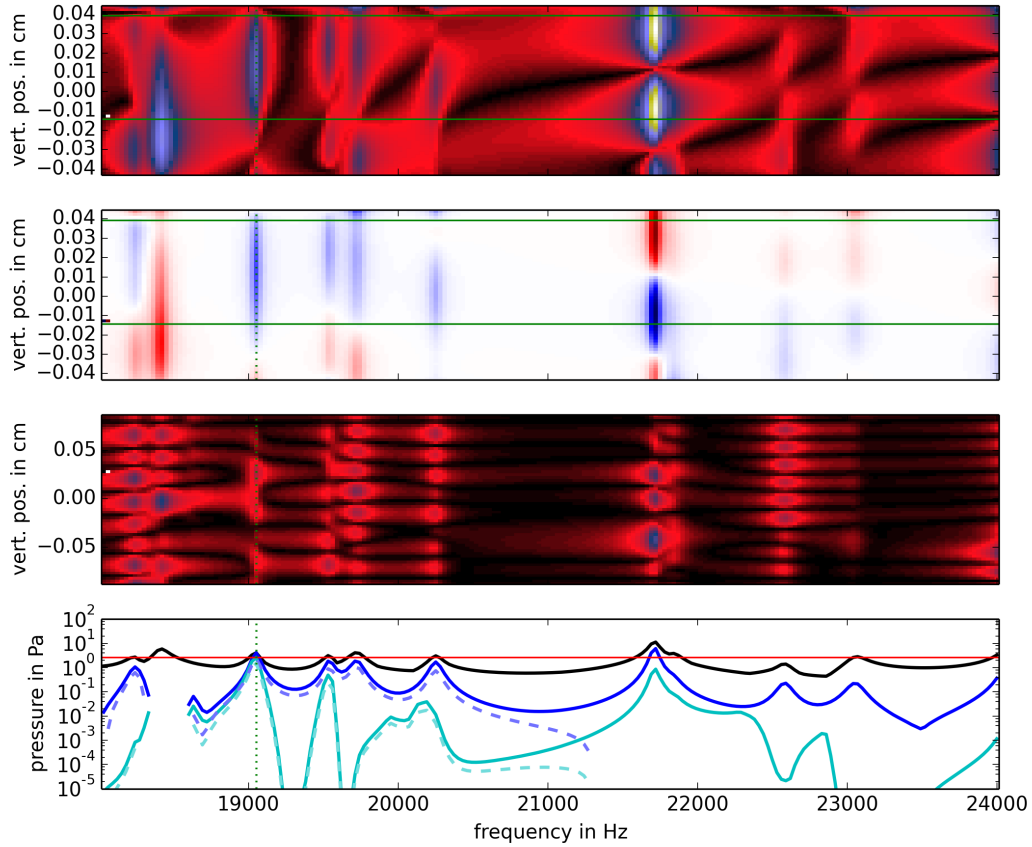


Figure 4.7 Score evaluation routine B: mode-discriminating version
This figure illustrates the score evaluation routine in a setup with mode shape discrimination geared at tuning towards the mode shape of the West-Howlett resonator design. The top plot contains the centreline pressure amplitude map and the region of interest excludes the lower part where the transducer is situated. The second colour map is the imaginary part of the same data in a colour scale where red stands for positive numbers and blue represents negative values. This allows discriminating between in-phase (red) and out-of-phase pressure peaks (blue) with respect to the transducer ring contraction. The third plot is the outer wall pressure map which among other interfaces contributes to p_{if}. The last plot shows the centreline pressure maximum (within the region of interest) in black and derived score functions further explained in the text.

The first measure is to restrict the transducer to off-centre positions in the lower half and to look for the centreline pressure peak not at the height of the transducer but only above it. This is indicated by the green lines bounding the region of interest in the first plot containing the centreline pressure map. The next step is to ignore in-phase pressure peaks and count only out-of-phase antinodes. An easy way to do it is to look at the imaginary part of the centreline pressure which is the second plot in figure 4.7 where red an blue regions designate positive and negative values. The search for the strongest out-of-phase peak means to look for the darkest blue in this map $\max(-\operatorname{Im}(p_{\text{centre}}))$, and this leads to the resonance shortly below 22 kHz. This is not an ideal mode shape, firstly because it has a high pressure amplitude on the upper piston front plate, and secondly with two and a half antinodes along the central axis each antinode is narrower and offers less space for bubble clusters than alternative modes. On top, it is not the sought-for West-Howlett mode shape.

Another postprocessing step can help to further narrow down the search, and this could be to multiply with the intended mode shape and integrate so that the scalar product is highest when the fit is best. What was in fact implemented was simply the multiplication of the centreline pressure between the green lines with a sine window before taking the maximum value. In the illustrative case shown in figure, this has the decisive effect which is to judge this resonator response not by the performance near 22 kHz but by the resonance at 19 kHz instead. The details are in the fourth plot at the bottom where the black line is the peak from the not windowed centreline pressure, the blue solid line corresponds to taking the maximum after multiplication with a sine window, and the solid cyan curve is the result after applying the wall pressure ratio-based penalty. There are additional dashed variants of the blue and cyan curves: they correspond to integrating along the vertical coordinate (within the region of interest) instead of simply taking the maximum. The result of the integral over $-\operatorname{Im}(p_{\text{centre}})$ times the sine window turns negative wherever there is more red than blue in the imaginary pressure map between the green lines leading to the curve breaks in the logarithmic plot. During the EAO trials the variant with "penalised maximum after windowing of $-\operatorname{Im}(p_{\text{centre}})$" yielded good results, so it was kept for most of the optimisation runs.

A note can be added that the exemplary frequency response plotted in figure 4.7 shows very nicely what was mentioned above, that first choosing the resonance and then calculating the penalty only at that point is a less efficient approach than calculating the penalised score across the whole frequency band first and then picking the resonance. With the simple approach the design variant would be judged by the performance of the 22 kHz resonance and given a mediocre score, whereas the latter approach allows discovering the good properties of the 19 kHz resonance and judging the candidate solution accordingly. The simple approach can even have the following strong side effect during evolutionary optimisation: it requires all other stronger mode shapes to be absent in the examined frequency band before the intended mode shape is counted, and this effect can act as a huge overall drag on the optimisation procedure.

4.2.5 The fitness function as a good example of discussing methodology and implementation

The resonators proposed below will probably not be the last design developments on the way to better SF experiments. These detailed descriptions are written in the spirit to promote EA optimisation as an efficient tool allowing to sustain a steady innovation process. Therefore, it is aimed at explaining particular implementations always in the context of both current and general goals. The fitness function with its penalisation scheme can serve as a good illustration. It would be a misinterpretation of the above to read it as a general recommendation that in the optimisation of acoustic resonators the objective function should always be purely pressure-based, and all other goals are better kept outside the blackbox optimisation procedure. It is absolutely not excluded that in future further efforts of improving the resonator designs, it might quickly become useful to employ completely different fitness and penalisation schemes. For example, it is imaginable to take one of the pressure-optimised designs discussed below and to further optimise it with respect to mechanical properties. One possibility would be to re-apply similar global or local search schemes with an objective function based on minimising stress peaks in structural materials while penalising only substantial setbacks in terms of $p_{\max}$ and r_{wp}. Another possibility would be the optimisation of subproblems, e. g. tuning a simulation of only a piston with the goals to maximise displacement, minimise stress peaks, and to reduce the internal bending of the front surface. Instead of ignoring the rest of the resonator a generic pressure BC taken from data of a good resonator could be loaded onto the piston front surface to make the subsystem simulation more realistic. In all such cases the same general thoughts and decisions about multi- versus single-objective, steep versus weak penalties or smooth versus unsteady forms or other implementations for combining or negotiating goals would have to be made. This is why increasing the detail level of this EAO case study and accompanying it with background and motivational thoughts is deemed to offer helpful information for future work on SF resonators or other applications of blackbox optimisers to real-world problems.

4.2.6 The sequence of one evaluation call

In order to evaluate one parameter combination corresponding to one candidate solution, the following sequence of steps has to be pursued:

1. **Input:** A new APDL input file has to be created by copying the master and modifying the parameters to be optimised.

2. **FE model:** The forced harmonic FE model is solved covering the whole frequency range of interest. The frequency stepping is coarse to reduce computational cost. Desired result data are written to text files.

3. **Objective function:** Output data are being read and postprocessed. The objective function is computed for each frequency step. The frequency where f_{obj} peaks is being identified.

4. **FE model:** Another FE simulation is carried out covering a narrow frequency band around this frequency in a sufficiently fine frequency resolution so meaningful peak values can be inferred.

5. **Objective function:** In the frequency axis close-up the choice of the frequency step to isolate and analyse can be made straightforwardly by simply identifying the centreline pressure peak. At that frequency step the wall pressure ratio is determined and the two values are used for the final fitness calculation via the penalty function.

4.2.7 The computational cost of the FEM-based EAO

With FE meshes of around 10^4 nodes, with the first scan of the wide frequency interval made in 100-200 steps and the second narrow scan in 60 steps, one design evaluation takes between one and three minutes on a dual-core laptop (Intel Core® 2 Duo, 2.16 GHz). The EAO runs were performed on an older Linux cluster of the KIT campus at similar speed. Considering the computational cost, parallelisation and the reduction of wasteful evaluation calls due to a crashing FEM software or postprocessing routine are of great interest. EAs are predestined for parallelisation because in each generation many candidate solutions can be evaluated independently. In the current case the bottleneck existed in the number of available user licenses for the FEM suite.[2] The issue of waste due to crashing evaluation calls deserves some consideration in detail.

4.3 Simulation robustness and foresight with the model parametrisation

Sometimes a chromosome cannot be properly evaluated because program errors happen. Either the FEM software crashes during preprocessing, solution, or postprocessing, or the external program for reading the data and condensing them into the fitness value runs into an error. The problem lies not particularly in how to treat these crashed trials in the EA, they can easily be kicked out of the gene pool by giving them a very bad score or even by making the algorithm aware of genomes to be ignored. The aspect creating more difficulties is the drag on the evolution if a fraction of the limited function evaluation budget is wasted on crashing simulations and the capacity of the gene pool to store and accumulate useful information is not used to its full potential. Therefore, the number of these crashes should be limited as much as possible, and in fact, a lot can be done in that direction by setting up the scripts and programs with foresight:

- by making sure that the model geometry parametrisation is set up in such a way that the searched solution space does not contain a lot of volume corresponding to geometrically infeasible designs,

[2] The licensing options for HPC computation tasks and design variation studies with Ansys have substantially evolved since.

- by preventing the possibility of geometries which are geometrically feasible and solvable by the FEM engine but would be unphysical in the real world,
- and by paying particularly strong attention to program code robustness.

The second point and examples of the EA optimiser maximising the fitness function by finding and improving unphysical solutions will be discussed in figure 5.16 when going through the history of EAO results. The first point, the issue of model parametrisation deserves the few sentences below.

4.3.1 Geometry complexity, parametrisation, and subspaces of infeasible solutions

As the complexity of a resonator geometry grows and ever more design parameters are opened up for the optimiser to be tuned, also the possibilities to make parameter combinations yielding infeasible designs can multiply. The first and most obvious question is then: why not favour simpler geometries? In the following, it is explained, why the simplest geometries will not always furnish hopeful starting scenarios for EA-tuned resonators, to what pitfalls the increased geometric complexity may lead during an EA optimisation, and how they can be addressed by carefully devising the right geometry parametrisation.

Examining some subtleties of the old piston geometry can lead to an understanding of the driving forces increasing the complexity of the resonator geometries. The structural parts of the resonator and the liquid volume are a system of masses and springs. Only by giving the optimiser enough degrees of freedom it becomes possible to explore many different ways of how all these masses and springs are coupled. To make the point with the eyes on an exemplary design detail, figure 4.8 compares a simple geometric model of the lower piston to a more complex alternative.

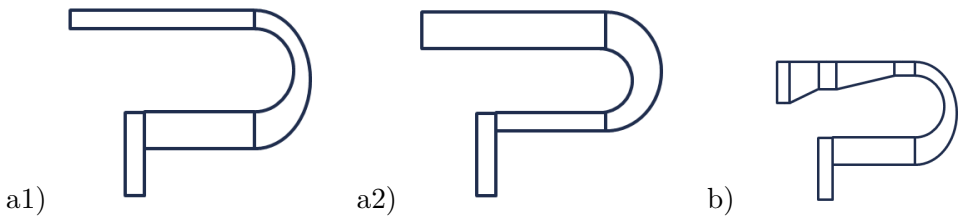


Figure 4.8 The drivers of geometry complexity
These sketches compare two versions of parametrisable piston cross section geometries, version (a) is rather simple and version (b) corresponds to a more complex segmentation. Setup variant (a1) exhibits a soft membrane-like front plate attached to a sturdy back, variant (a2) may allow a rather rigid front plate to vibrate against the holding tube. A parametrisation of geometry (b) can allow for many more degrees of freedom of tunable masses on the front plate, the side, or the back of the piston connected either sturdily or via thin parts serving as tunable soft springs.

In the West-Howlett design the piston is held by the elastic silicone connection sealing the gap between piston tube and the outer glass hull. The main task of the piston is to furnish a boundary to the central fluid domain which is neither a zero displacement nor a zero pressure boundary. At the same time it is also a

boundary to the fluid volume behind the piston. The front surface of the piston oscillates back and forth, and it has to have the right mass and spring behind it, in order to allow the liquid bulk next to it to adopt the desired vibration mode at the proper frequency. Of the simple geometry version, two setups are shown. One (a1) with a thin and soft oscillating front plate being held by a sturdy piston back part and a different setup (a2) where the front plate and rim are sturdier and where the piston back plate is the part predestined for most of the elastic bending. This would allow the whole piston front including rim to oscillate against the rest. The point is that the piston can fulfil its main task only if it functions as part of the whole resonator. The piston design cannot solely be enslaved to furnish the right acoustic impedance for the fluid volume in contact with its front plate, it must have enough design degrees of freedom to fit into a vibration mode where it is coupled in between the fluid subvolumes in front and behind the piston and with the other structural parts like an end plate or the glass hull directly. The geometry depicted on the right exemplarily shows how more DOF can be introduced into the piston design in a way such that more internal flexure modes become possible. Sufficient DOF are crucial for allowing the piston to serve as beneficial BC for both at the same time: the liquid in front of it and the structure connected to its back side.

With increased geometric complexity and a higher number of DOF open to tuning, the options of how to parametrise the model multiply. The search domain seen by the optimiser is a cuboid if for each parameter x_i there is an allowed interval $[a_i, b_i]$. When making the choice between different model parametrisations, there are two very important criteria to pay attention to: how large is the volume fraction of the search domain representing solutions which are geometrically not feasible, and what shape does the remaining volume of feasible solutions take? Figure 4.9 compares two possible parametrisations. One is an example for minimising the volume fraction of infeasible solutions and thus the probability of futile and wasteful evaluation trials. This is the reason why the choices of tuning parameters listed when discussing the new designs in the following chapter (e. g. fig. 5.26) may seem unintuitive and complicated at first.

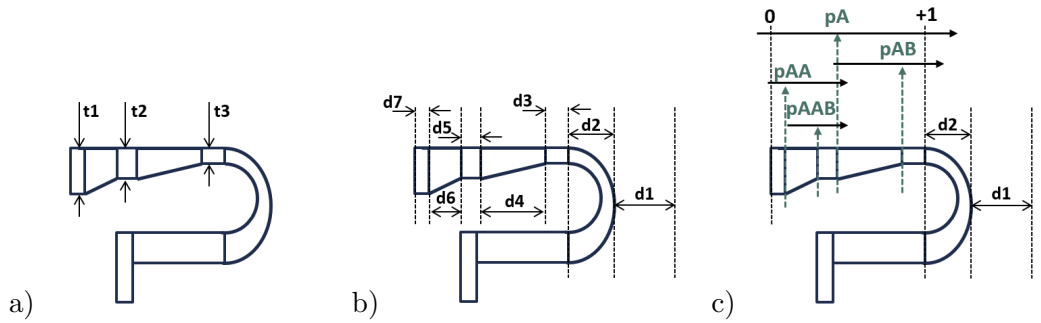


Figure 4.9 Geometry parametrisation and search efficiency
The diagrams illustrate parametrisation variants for a complex piston geometry. The thickness parameters in drawing (a) can be chosen independently. The horizontal distances in diagram (b) have to abide to the constraint of fitting inside a resonator of given radius r. Diagram (c) symbolically shows a changed parametrisation devised for achieving only feasible geometries and for being useful in an evolutionary scheme of parameter inheritance and mutation.

In the exemplary illustration in figure 4.9 the piston front plate is divided into

segments with three independently tunable thicknesses and trapezoidal transition regions. On the left (a), the three thickness parameters (t_1, t_2, t_3) are indicated. Sketch (b) in the centre shows the seven horizontal distance parameters $(d_1, d_2, ..., d_7)$ arising from the segmentation. d_1 is the gap between the piston rim and the inner radius r of the cylindrical main glass wall. The size of r will be considered as a given boundary condition for this subproblem. It is either fixed or one of the principal tuning parameters. If the seven distances d_i have to add up to r, then there are six remaining DOF. What is the best parametrisation for them? The simplest option is to leave the parametrisation as is and to turn all the d_i except one d_j into tuning parameters x_i for the EA, d_j becomes dependent so it covers the remaining space. The problem with this parametrisation is that many new trial chromosomes $\vec{x}$ will translate into infeasible geometries where $\sum_i d_i > r$ which would result in a negative d_j.

The schematic on the right symbolises a different parametrisation. Here, d_1 and d_2 stay the same as before, but the remaining parameters are transformed to be not absolute distances any more but subdivisions of larger into smaller segments by fraction. These divisions are symbolised by the dashed vertical arrows sliding along the solid horizontal arrows. The longest dashed arrow with the parameter label pA slides along the top horizontal arrow symbolising the corresponding coordinate system. pA divides the piston front plate into two segments, and the parameter value is the fraction of the left section. pAA and pAB are analogous further subdivisions of these two segments. The parameter $pAAB$, finally, covers the last remaining DOF and opens up the next layer of the subdivision branching structure. The optimiser can be given the allowable interval $[0, 1]$ for each of these parameters (however, something like $[0.1, 0.9]$ will be more practical in terms of allowing any geometry building and meshing script to be more robust and concise).

By going from absolute distances to fractional subdivisions this way, the search engine can have the freedom to explore the whole design space without any chance to hit on infeasible solutions of the geometric subproblem. For the remaining absolute distance parameters d_1, d_2, and eventually r, it is now easy to define reasonable boundaries negotiating design freedom with no or small probabilities of infeasible solutions. Of course, they can also be turned into fractional subdivision parameters if need be.

Furthermore, it can be assumed that the transition from distances to proportion parameters will in most cases be associated with a beneficial influence on the topology of the search landscape. All masses, springs, and fluid volume sizes are interdependent in the resonator, and one part must be dimensioned in proportion to the other. As a thought experiment one can think of a rectangle which can be parametrised either by width and height or alternatively by size and aspect ratio. The first parametrisation will have solutions of similar proportion forming a thin region stretching diagonally through the parameter space. In the alternative parametrisation, any such subregion can be found by tuning only the aspect ratio and ignoring the size parameter, turning a 2D into a 1D problem. For resonators size always corresponds to frequency. Keeping it separated from issues of proportioning may in many cases simplify the situation.

An additional point, which should be kept in mind, is that, unless no special

modifications are made to the search algorithm, the failed evaluations entering with a default bad score, still contribute to the flow of information processed by the algorithm and influence its state. The corresponding regions appear to the searcher as plateaus. This is why the shape of the subspace representing the feasible solutions is of great importance. If the function minimiser sees a large plateau covering most of the search domain, just broken up by a long narrow canyon, and all the subspace of feasible solutions with all its local maxima and minima, with all its structure and complexity is squeezed into this narrow canyon, then most EAs and many other search algorithms will not be able to adapt to the shape of the canyon, with CMA-ES and the downhill-simplex algorithm being notable exceptions. So, besides the direct waste through failing evaluation calls, the wrong geometry parametrisation may induce, depending on the algorithm, a much larger drag on the search efficiency arising from a disadvantageous overall shape of the resulting search landscape.

4.4 More visual diagnostics: stalling local simplex search

The preceding paragraphs described some issues of how to prepare the engineering problem of resonator tuning to turn it into an application case where evolutionary global search can be applied usefully and efficiently. It means looking at the objective function from the perspective of the searcher. The same is important to ensure effective application of a local search (LS) routine. Each local search algorithm has its own properties, and it should be checked that there are no features in the objective function which based on the local searcher's paradigms and implementation would lead to inefficiency or even stall the search. Figure 4.10 illustrates what it means for the deterministic *downhill-simplex* search algorithm (*Nelder-Mead* algorithm [323]) if the objective function is not smooth nor steady due to numerical and/or discretisation effects.

The diagrams in the figure show the working principle of the downhill-simplex algorithm and how it can fail during the FE model tuning task. The basic idea of this LS technique is to efficiently pursue a gradient-following path by reflecting, expanding, and shrinking a simplex. The basic operations of the used algorithm implementation are sketched in the upper left drawing for the 2D case where the simplex is a triangle formed by the black-circled points labelled $x1$, $x2$, and $x3$. Abiding for now to the "downhill" paradigm of function minimisation the point $x3$ is the worst vertex of the simplex representing the highest objective function value. Picture (a) shows the first type of operation tried to replace $x3$ which is evaluating the reflected point labelled xr constructed by mirroring the worst point across the mean vector of all other simplex corners. If this heuristic yields success and xr indeed turns out to be the new best point, then the expansion of the simplex into the same direction is tried, depicted in sketch (b) where xer represents the result of the extended reflection. In case xr does not turn out to be the new best points the algorithm tries two more conservative operations, the contraction of the reflection point yielding xc and the inside contraction (c) yielding xcc, a simple retreat from the worst point. A very expensive operation of last resort is depicted in sketch (d): it is the shrinking of the whole simplex by pulling back every vertex except $x1$ towards $x1$. The heuristics of the downhill-simplex algorithm are intended for the situation

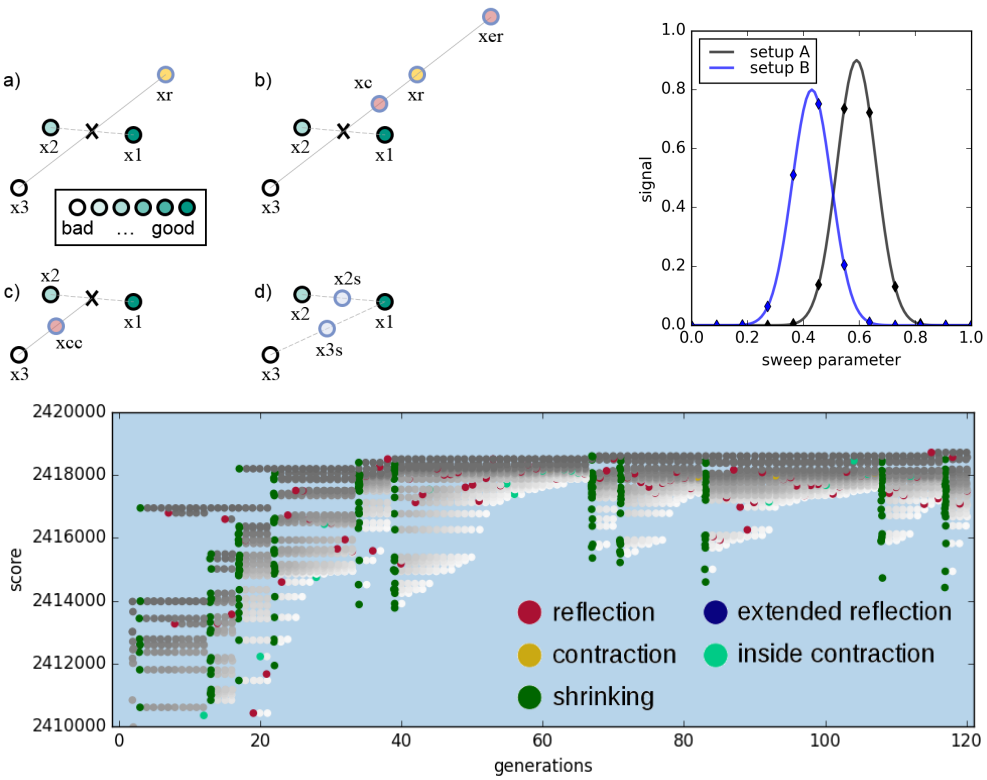


Figure 4.10 Visual diagnostics E: downhill-simplex requires a smooth function
The sketches in the upper left show the basic operations of the Nelder-Mead algorithm [323]. The three black-rimmed circles labelled $x1$, $x2$, and $x3$ designate the current simplex in a 2D search space. The four test points which can be generated are the reflection point xr shown in (a), extended reflection xer and contraction xc in (b), and inside contraction xcc in (c). The last branch in the algorithm's logic structure is to shrink the simplex, shown in (d). The diagram in the upper right shows the effect an axis discretisation has on a peak measurement. In the employed resonator FEM simulations the frequency axis as well as the spatial mesh coordinates are discretised. The score history plot in the lower half of the figure shows the end of a downhill-simplex search after the size of the simplex has shrunken so the scale of fine-structure on the objective function caused by discretisation effects becomes the dominating effect. This landscape is misleading for the simplex search resulting in frequent shrinking. That the shrinking operations keep worsening the scores proves that the search landscape is not smooth.

when the structure scale of f_{obj} is larger than the simplex size. If this is not the case, it is intended that the desirable situation is enforced by shrinking the simplex to a size smaller than the structure on f_{obj}.

The plot in the upper right of figure 4.10 shows two arbitrary functions with gaussian peaks of same width but different height. We can imagine them to be representative of scans of a sound pressure peak, either on the frequency axis or along the spatial coordinate z. When evaluating resonator FEM simulations, the sound pressure profile is being probed on a discrete grid both in the frequency and the spatial domain. In the diagram a particularly coarse probing grid is depicted by the diamond symbols. It can be seen that judging by the grid data the left peak would be inaccurately determined as the higher peak. The limited accuracy can be understood as noise or better as a fine structure added to the peak height signal.

The downhill-simplex algorithm can follow a gradient properly only for so long as the level of noise on the objective function is of much smaller amplitude as the

difference in signal gathered from the simplex corners. The lower plot shows a score history diagram illustrating how the downhill-simplex algorithm stalls early on after having achieved only little improvement. A characteristic feature are the repeated vertical columns of green dots representing the occurrence of the simplex shrinking operation. It is noteworthy that pulling the corners closer towards the best point does not lead to improvement but consistently worsens the score values. This is a clear indicator that the simplex does not shrink into a local valley in a smooth surface. The explanation here is that the simplex search stalls in the rough fine-structure of the objective function due to the numerical effects of the FEM simulation and the postprocessing. It is perhaps also noteworthy that the score history visualisation of the deterministic downhill-simplex algorithm offers a clear hint towards the problem whereas score data from stochastic algorithms would most probably hide it well.[3]

During the EA+LS optimisation case studies presented in the next chapter it was ensured by a fine FE mesh resolution of 1.5 mm and a fine frequency stepping for the close-up scan of $\leq$ 4 Hz that the numerical noise level on the evaluated objective function was low enough to make premature stalling of the local simplex search improbable. For future optimisation trials it might nevertheless be important to know that there are straightforward measures available to drastically reduce the amplitude of the numerical noise (without increasing the resolution of the FEM simulations) simply by smoothing through curve fitting. Quadratic or spline fits may very well represent a workable physics-agnostic first option. Better yet would be physically motivated fitting functions, e. g. gaussian peaks on the frequency axis and sinusoidal curves for the sound pressure mode shapes (the vicinity of antinodes). An even better way of a fit-based measurement of mode shape peak heights might be done via decomposition into spherical harmonics [190] because this would allow taking into account a 2D or 3D field of data surrounding a pressure antinode.

[3]This is a motivational thought for Glover et al. for favouring deterministic over stochastic algorithm features whenever possible [172] (see also appendix section T.4.8).

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
d_i	distance parameters (exemplary parametrisation display)
f_{obj}	objective function
f_{FD}	Fermi-Dirac distribution function
f_{pen}	penalty function, contribution to an objective function
$\tilde{f}_{\text{pen}}$	transformed penalty function projecting from $\mathbb{R}$ onto $[0, 1]$
p	sound pressure amplitude
p_{centre}	sound pressure maximum along the centreline (radial axis)
p_{if}	sound pressure maximum throughout the fluid-structure interface
$p_{\text{lim}}, p_{\text{lim},1}, ..$	sound pressure limit/threshold values
p_{max}	sound pressure maximum throughout the resonator
$pA, pAB, ..$	proportion parameters (exemplary parametrisation display)
Q	quality ("pointedness" of a resonance peak)
$\mathbb{R}$	real numbers
r	radius
r_{wp}	wall pressure ratio, a measure for judging resonator design quality
t_i	thickness parameters (exemplary parametrisation display)
x_i	design parameters
$\vec{x}$	solution candidate, chromosome, search point
z	axial coordinate

List of Greek quantity symbols

Symbol	Description
α_1, α_2	generic scaling factors
β, β_1, β_2	parameter for transition width of a Fermi-Dirac distribution curve
θ, θ_i	penalty function threshold parameters
μ	anchor/transition point of a Fermi-Dirac distribution curve

List of abbreviations

Abbreviation	Description
APDL	Ansys Parametric Design Language
BC	boundary condition
CMA-ES	evolution strategy with covariance matrix adaptation
DOF	degree of freedom
EA	evolutionary algorithm
EAO	evolutionary algorithm optimisation (meaning optimisation by evolutionary algorithm)
FE,FEM	finite element (method)
HPC	high-performance computing
KIT	Karlsruhe Institute of Technology (Karlsruher Institut für Technologie)
LS	local search
MO,MOO	multi-objective (optimisation)
SF	sonofusion
SO,SOO	single-objective (optimisation)

Chapter 5

EA-optimised SF resonator design proposals

This chapter contains the presentation and discussion of the main results of this thesis: optimised resonator geometries obtained by applying evolutionary algorithm optimisation (EAO) to FE models. Three new resonator geometries were investigated. They were optimised using CMA-ES and THEA. Based on different model symmetries and parametrisations different optimal design setups are yielded. The objective function is a quality measure derived after the application of various data post-processing subroutines. It is shown that a thoughtful design of the entire evaluation routine is on the one hand a crucial determinant of successful optimiser application and on the other hand it offers important degrees of freedom for steering optimisation outcomes. A geometry instance of the West-Howlett resonator design used by Taleyarkhan et al. for SF trials was optimised in order to have a benchmarking baseline for the resonator performance.

5.1 Chapter framework

5.1.1 Chapter structure

This chapter is framed by an introductory text on data presentation conventions and FE model settings and a concluding text section summarising and discussing the SF resonator EAO results. The central part is dedicated to the EAO case studies. The four investigated geometries are discussed in separate sections starting with little introductory texts with the main content coming in a series of large figures with discussions in the caption texts. This offers a guiding structure of going from case study to case study with intermittent highlights on special aspects.

5.1.2 Figure legend for EAO result plots

All optimisation results on the four examined geometries will be presented in standardised figures to make them easily comparable. The figures contain four different plot types and two tables. The plot types show the geometry, the pressure field amplitude, a pressure field snapshot, and a snapshot of the deformed FE mesh.

The **geometry plot** is a snapshot of what in Ansys is called an area plot (APDL command: `aplot`) showing the axis-symmetric 2D geometry as a patchwork of areas to be meshed one after the other. The building blocks are mostly rectangles, circle segments, and derivatives via operations like area addition, subtraction, intersection, merging, or stretching. The tuning parameters enter the FE model by directly influencing the shape of these areas. The colour code of the area plot mirrors the material attribution to be used in the FE mesh creation routine. The colour legend is shown in figure 5.1.

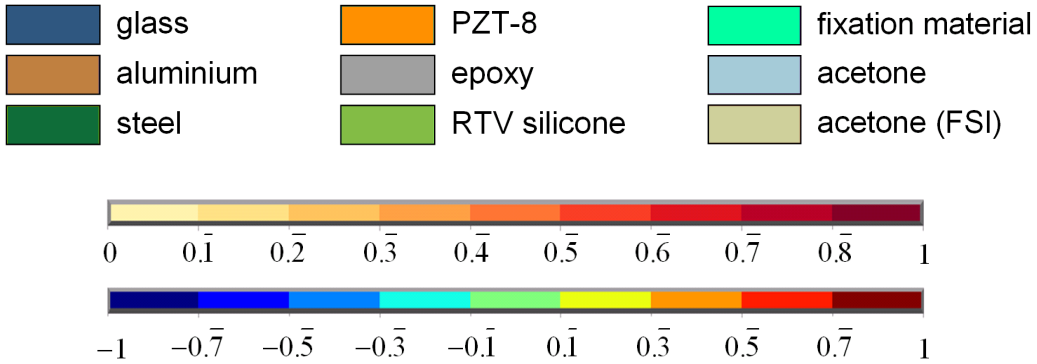


Figure 5.1 Colour legend for EAO result plots
The upper part of the figure contains the colour code legend for the material attribution in the FE geometry and mesh plots. Two colours are necessary for the acoustic fluid domain in order to distinguish internal finite elements where pressure is the only nodal degree of freedom (DOF) from boundary elements with additional nodal displacement DOF enabling fluid-structure interaction (FSI). Below, the colour scales of the contour plots are given. The axes are given in fractions of $p_{\max}$. The upper scale covering the interval $[0, 1]$ is taken for pressure amplitude plots, the lower one covering $[-1, 1]$ applies to snapshots of the pressure/tension field.

A **sound pressure amplitude plot** is shown in a reddish colour scale next to a **sound pressure snapshot plot** in the common "jet" colour scale ranging from blue (pressure below the reference pressure p_0 or state of tension) to red ($p > p_0$). These colour scales, illustrated in figure 5.1, are always normalised to the maximum occurring pressure amplitude. This means they cover the intervals $[0, p_{\max}]$ and $[-p_{\max}, p_{\max}]$ which explains why the reddish colour scale is exhausted in its full range in each plot while for the jet scale this is often not the case. The colour scale normalisation allows to obtain a maximum of topology information from each plotted pressure field, but in comparisons of different resonators it hides the information about which one achieves the better sound pressure performance. These values are to be found in the tables below the plots.

The last one in the row of cross section snapshots is the **deformed mesh plot**. This type of plot shows very well how the FSI layer of acoustic fluid elements connects the rest of the acoustic fluid domain with the structure in motion. The plots were all made with a displacement amplification factor of 800.

The two tables below each case study figure contain important characteristics making the different cases comparable. On the left, a **resonator specs table** lists the key features of physical nature of the optimised solution: the resonance frequency f_{res}, the maximally achieved sound pressure p_{max}, the Q-factor, and the wall pressure ratio r_{wp}. The computation of r_{wp} takes into account all of the fluid-structure interface. Interface pressures at some key locations are printed separately: p_{upc} is the sound pressure amplitude measured at the centre of the upper piston front plate, p_{lpc} is the counterpart value from the lower piston, p_{p} is the maximum sound pressure found across the whole surfaces of both pistons, and p_{wall} is the maximum sound pressure along the inside of the cylindrical main glass wall. In some cases, displacement amplitudes, denoted with u, and displacement amplification factors, denoted with A, are indicated.

After the decision on the frequency step to be examined, all these sound pressure values are taken right at this frequency. As described before, the EAO programs went through many development steps. The way how the frequency decision and the fitness computation were made during the EAO runs looked very different from time to time. Therefore, a re-analysis of the optimised FE models has been carried out in order to yield the comparable values in the specs tables. Just the close-up of the resonance peak chosen by the optimisation routine was re-examined. It involved identifying the frequency step yielding the highest p_{max} in the liquid in the region of interest along the central axis as the resonance frequency f_{res}. Subsequently, the listed pressure amplitudes and r_{wp} were determined. How the optimiser evaluated and scored the solution candidates at the time is indicated in the optimisation details tables and case descriptions.

The table with **optimisation key facts** on the right will similarly achieve comparability at a glance in terms of optimisation approach taken in the respective cases. It indicates used optimisation algorithms, population sizes, generation or iteration numbers, the spent amount of evaluation calls, and not least the type of evaluation routine (objective function) used.

Each figure caption contains a **list of optimised parameters** helping to compare the optimisation runs by the number and type of free tuning parameters and implicitly indicating the degree of allowed asymmetry between upper and lower halves of the resonator geometries. Listings of the final parameter values yielded by the optimisation runs would appear as lengthy rows of mostly not very helpful numbers and are therefore given only for relevant selected cases in appendix W.

5.1.3 Common FEM simulation settings

The discussed optimisation case studies build on SF resonator FE models established as part of preceding research projects at KIT and RPI. The modelling approach and some results were published in conference papers [435–437]. As the developed resonator models are the basis of the resonator optimisation runs discussed below, a corresponding appendix chapter containing detailed model descriptions is added in appendix Q. In particular, material data are listed in tables Q.1-Q.3. The FE modelling suite ANSYS® (Mechanical, APDL) was used for implementing the 2D axis-symmetric resonator FE models.

In order to keep the different FE models for optimisation comparable, the element size and the damping settings were kept the same for all EAO cases. Instead of material-specific damping parameters a damping ratio of $\zeta = 0.003$ was set applying globally to all structural[1] finite elements of the model (APDL command: `dmprat,0.003`). In high-Q systems, damping has a heavy impact on the heights of resonance peaks. If the single materials have very different damping constants, then a major goal is introduced into the geometry optimisation which is to minimise bending in thin parts of strongly damped materials and concentrate the elastic energy storage in other parts of the geometry, ideally only in the materials with lower damping. As long as the data on damping properties of the used materials are not so reliable, it has been decided to keep this aspect away from interfering with the optimisation task during this attempt of EA application.

In terms of element size, the APDL command `aesize,all,1.5e-3` was used to globally set a maximum element edge length of 1.5 mm. Local mesh refinings were not considered due to the risk of making the APDL scripts more complex, less robust, and the resulting models less comparable under large parameter variations.

The material properties were the following: the structural materials were the ones of table Q.3; for glass the dataset labelled "glass-1" was taken, for epoxy the set "epoxy-2". The liquid was acetone with the properties of table Q.2. For the piezoelectric ceramic the "preliminary" set of constants labelled "p8d" in table Q.1 was applied. Consistency throughout all optimisation cases was considered a high priority, so the PZT-8 material data was deliberately never updated to a more realistic set during EAO trials.

Generally, the FE models were used for frequency sweeps of forced-harmonic analyses, whereby a voltage drop of 100 V between inner and outer surfaces of the transducer ring served as the boundary condition for excitation of the oscillatory motion response. The liquid-filled inner part of a resonator was modelled as purely acoustic domain – this means pressure is the single nodal degree of freedom (DOF) – which entails the necessity of activating a dedicated *fluid-structure interaction (FSI)* formalism (APDL command `sf,,fsi`) for translating between displacement and force on the side of the structural domain and the pressure DOF on the acoustic side.

5.2 The West-Howlett geometry

The preceding chapters described a few subtleties which needed to be resolved in order to find the right EA for the resonator tuning task and to interface it suitably with the simulation. Complications need motivation to be surmounted, and sometimes it comes in the form of a demonstration experiment with convincing effect. The German language offers the nice word "Aha-Erlebnis" for such an eye-opener with "aha" effect. For the current context, the proof of concept, the demonstration that the EA approach is worth the effort, is revealed by the comparison of two optimisation results. The first is presented in figure 5.3 showing the result of the long process of manual tuning aimed at maximising the sound pressure achieved

[1] The used acoustic fluid elements do not allow for simulating material damping.

in the FEM simulation of the West-Howlett-style resonator (i. e. the conventional SF resonator geometry used by Taleyarkhan et al. and other researchers conducting replication experiments). With a driving voltage of 100 V it creates a sound pressure amplitude of 22 bar. The other geometry setup, depicted in figure 5.4, is the result of applying the THEA algorithm in a simple version to the simulation. Right upon the first four attempts pressure amplitudes of 20-27 bar were reached. The preceding figure 5.2 introduces the parametrised geometry.

The West-Howlett resonator geometry had been tuned manually in order to find a suitable starting point for the preceding sensitivity study (see apdx. Q) which ultimately triggered the presented SF resonator EAO study. The manual tuning process could build on previous experience with the FE model of that particular geometry and the result of a preliminary manual optimum search, and it had been conducted in the following way: The parameters determining the shapes of the main liquid volume and the glass pistons had already been identified as the decisive ones. Next, the six strongly influential parameters were varied one after the other which led to the identification of three promising setups. They were starting points for two further steps of systematic search by scanning 2D slices of the whole parameter space. The result had been the setup depicted in figure 5.3 which delivers 22.04 bar in the simulation. Between two and three thousand partially scripted FEM calculation jobs were consumed by this search.

When the hybrid EA was first tried on this FE model, nine parameters were tuned in four short EA runs over 20 to 40 generations. The chromosome populations tended towards different local optima, in one case with a single-antinode pressure profile (20.6 bar), in the other three cases towards asymmetric profiles with two antinodes (25.7, 26.0, and 27.2 bar). The best case is shown in figure 5.4. It can be seen that the EA optimisation introduced no problematic aspects into the sound pressure field (like particularly high amplitudes near walls) even though the evaluation was simply according to $p_{\max}$ (lacking any fitness function sophistication in the form of wall pressure-based penalty schemes as discussed in the preceding chapter). These global[2] EA searches consumed around 8000 evaluations.

Comparing the computational cost, the EA approach seems much less efficient. But this is not really what counts. The more important aspect is the time and effort the human researcher needs to put into the task. Once the program code and infrastructure for EA optimisation has been put in place, starting four optimisation jobs and later analysing their results is a surprisingly little amount of work compared to many days of repeated data analysis, thinking, and renewed scripting needed permanently during the manual exploration of the design space. The substantial

[2]Generally, the justification for a global search approach in terms of performance and suitability can and should be scrutinised by systematic comparisons with local search trials. That local downhill-simplex search is not capable of reliably yielding highly optimised resonator setups is shown in the section on geometry A, particularly with figure 5.19. Concerning the West-Howlett geometry, later conducted local search trials showed that, depending on the parametrisation and the willingness to accept high wall pressure ratios, a substantial further improvement of the depicted solutions in terms of $p_{\max}$ can still be possible. These unsystematic trials are however not discussed here, leaving the focus of this section on the West-Howlett geometry on the results which had yielded a proof of concept of the EA-based approach and which had motivated the decision in favour of the EAO-based investigation of new geometries at the time.

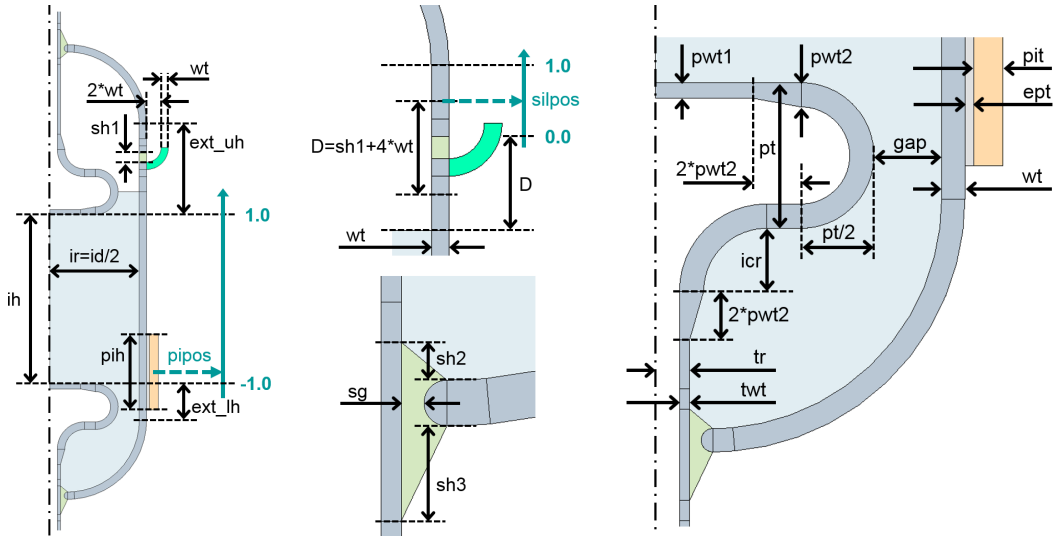


Figure 5.2 The parametrised West-Howlett geometry

The drawings show details of the 2D-axis-symmetric geometry of the FE model of the West-Howlett-style resonator. Parameter descriptions are given in the table below. Values are given in the last column for dimensions kept constant during optimisation and ranges for others. Values are in millimetre unless they belong to one of the dimensionless control parameters indicated with green arrows above. In the drawings proportions were changed in several places in order to make details better visible. The drawings together with the table listing represent the complete parametrisation necessary to describe geometry variations unambiguously. This implies that all round shapes are circular, not elliptic. The quarter circle (depicted in cyan) attached to the upper rim of the main glass wall is made of the artificial "weak spring material" (see table Q.3, p. 425) used to fixate the FE mesh against translation. A zero displacement BC applies at its upper edge. Furthermore, a parameter naming convention is introduced whereby the suffixes "uh" and "lh" are used to designate parameters with different values in the upper and the lower half of the resonator geometry.

label	description	value
id	inner diameter of main volume	59.2
ih	inner height of main volume	[80;140]
pit	piezo ring thickness	3
pih	piezo ring height	25
pipos	parameter determining the transducer's vertical position	[-1;1]
ept	epoxy layer thickness	0.5
ext_uh	extension from reflector to cap base (upper half)	[20;60]
ext_lh	extension from reflector to cap base (lower half)	[15;40]
wt	wall thickness of main cylindrical glass wall (hull)	[1;4]
sh1	height of silicone bead sealing the top head	3
silpos	silicone bead position	0.5
sh2	silicone bead height (inside)	1
sh3	silicone bead height (outside)	3
sg	silicone gap around piston holding tube	0.5
gap	gap between piston rim and inner surface of glass hull	[4;12]
pt	piston thickness	[10;30]
pwt1	piston wall thickness 1 (front/reflector plate)	[1;4]
pwt2	piston wall thickness 2 (rim and back side)	[1;6]
tr	radius of piston holding tube	3.55
twt	tube wall thickness	1.05
icr	inner curvature radius (connection piston-tube)	6.5
tl	tube length (measured from piston back side)	100

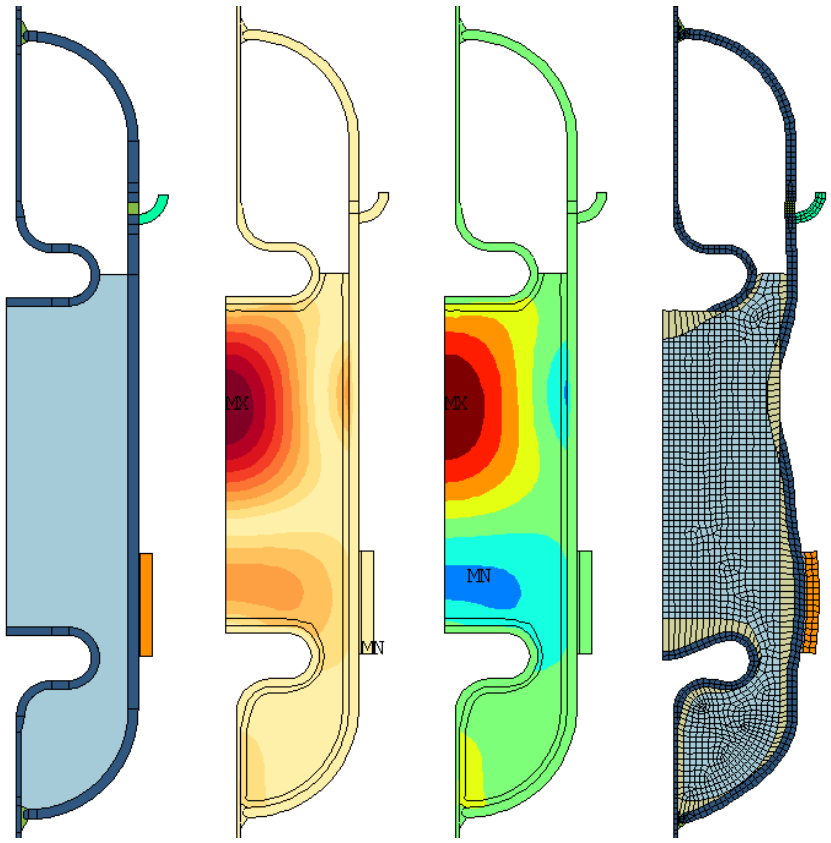


Figure 5.3 West-Howlett design: manual tuning
The resonator model of the West-Howlett geometry was optimised before it became the subject of the sensitivity study described in the appendix chapter Q.3.1. The final result of the manual procedure of variation and improvement is depicted above. The procedure involved testing parameter sensitivities, going to thicker piston front plates (2 instead of 1 mm) and subsequently first scanning the parameters $ih \times gap$ on a 7×19 sample grid then $ih \times pipos$ in 7×10 steps. This resonator setup can reach a pressure of $p_{\max} = 24.85$ bar if the set of material-dependent and partially calibrated damping constants described in appendix Q.3.1 are used. After switching to a global damping ratio of $\zeta = 0.003$ (i. e. harmonisation with the other EAO cases) this reduces to 22.04 bar.

The peak internal pressure amplitude of 22.04 bar has to be put in relation with the maximum pressure amplitude occurring along the whole fluid-structure interface. In this setup the wall pressure peak occurs not on the pistons but on the inside of the main glass wall where it exhibits the largest radial displacement, approximately on the same height as the strong pressure antinode on the central axis. This wall pressure peak has an amplitude of 8.14 bar corresponding to a wall pressure ratio of $r_{\mathrm{wp}} = 37\,\%$. By consequence, a small region of blue colour appears in the pressure field snapshot indicating amplitudes beyond 33.3 % on the tension side. By investigating how far the frequency has to be detuned so the centreline pressure peak decays in amplitude to a fraction of $1/\sqrt{2}$ a mechanical Q-factor of 389 can be deduced (i. e. peak width at half dissipation power, see table J.2 or eq. K.34).

The optimisation details table below indicates approximately the simulation budget which was necessary to accomplish the systematic last part of the optimisation history having led to this particular optimisation result, but it does not account for the establishment of the pre-existing knowledge base upon which the manual search could build.

(list of tuning parameters (latest phase): *ih*, *pipos*, *gap*, *pwt1*)

resonator specs	
f_{res}	21 213 Hz
$p_{\max}$	22.04 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	2.06, 1.29, 8.14 bar
p_{wall}	8.14 bar
r_{wp}	0.37
Q_{mech}	389

optimisation key facts	
method	human systematic search
scanned	19.5-21.5 kHz
f_{obj}	increase $p_{\max}$, keep r_{wp} low
N_{eval}	2500

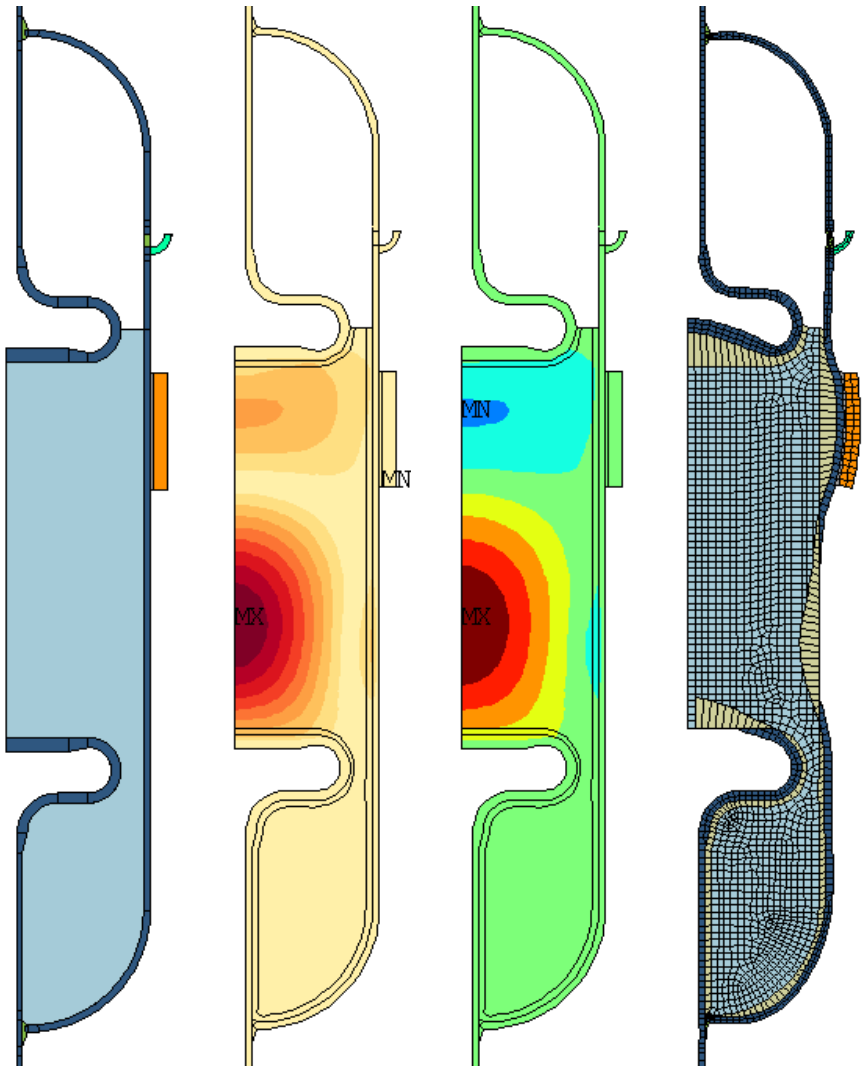


Figure 5.4 West-Howlett design: EA tuning
This figure shows the best result of four short experimental EAO runs on the West-Howlett resonator geometry. There was no pre-optimised starting point for the EA to build on. All starting populations were random-initialised. The allowed intervals for each of the nine free parameters was the only guideline. Like in the rest of the EAO cases, the FE model was damped with a globally applying damping ratio of 0.003. The depicted solution, the best-scoring chromosome in the final population, has a resonance at 20 448 Hz with $p_{\max} = 27.21$ bar. With $r_{\mathrm{wp}} = 25\,\%$ it exhibits a wall pressure ratio which is quite low compared to the manually optimised setup. This particular EA run consumed 1600 evaluation calls.
(list of tuning parameters: *wt*, *gap*, *ih*, *pt*, *ext_uh*, *ext_lh*, *pipos*, *pwt1*, *pwt2*)

resonator specs	
f_{res}	20 448 Hz
$p_{\max}$	27.21 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	2.47, 6.62, 6.72 bar
p_{wall}	6.62 bar
r_{wp}	0.25
Q_{mech}	468

optimisation key facts	
EA	THEA
scanned	19.5-21.5 kHz
f_{obj}	$p_{\max}$
$N_{\mathrm{population}}$	80
generations	20
N_{eval}	1600

reduction in human work, the achieved good solutions with no drawbacks, and the expected speed-up of the envisaged iteration cycle of simulating and judging many variants of new resonator designs, therein lied the "aha"-experience which motivated the rest of the EA-based optimisation work.

The West-Howlett resonator and wall displacement amplification

The history of the resonators manufactured at RPI (see fig. I.5, p. 292) shows that several times a potential resonator design improvement was sought in deviating from the West-Howlett design by going to a symmetric setup with the transducer in the central plane. The background thought may have been to put the driving power closer to the location where it is needed. The simulations carried out later as part of the RPI-KIT collaboration [435] yielded a possible explanation of wherein consists the advantage of the off-centre position of the transducer: it allows for the leveraging of a mechanical displacement amplification effect through the mode shape of the outer glass hull. The simulations of tuned West-Howlett resonator setups consistently show that the radial displacement amplitude of the cylindrical glass wall a few centimetres above the transducer is larger than directly at the height where it is mounted, where glass and piezoceramic are glued together with epoxy and have to move in lockstep. The point of largest displacement amplitude in the glass is at a similar height as the intended cavitation site, the largest pressure antinode in the centre. The transducer excites the motion at the location of the lower antinode where the sound pressure rise is in phase with the transducer contraction. The cavitation site is in the upper antinode, where both the pressure and the glass wall contraction are out of phase with respect to the transducer. These facts allow the interpretation that for a given transducer and a given voltage limit (imposed by the peripheric electronics and/or the piezoelectric material itself) a resonator can be optimised if it enables the development of vibration modes involving a mechanical amplification of the transducer-given displacement amplitude, and if structural parts with enlarged motion amplitudes are bounding the strongest sound pressure antinode representing the intended cavitation site.

Postprocessing the FEM simulation output, the displacement amplification can be easily computed. In order to have a comparison baseline when discussing the new designs, the numbers for the purely manually tuned resonator setup shown in figure 5.3 are given. This resonator setup features a maximal radial glass wall displacement at the height covered by the transducer of 3.6 µm. The out-of-phase displacement peak five centimetres above has an amplitude of 7.5 µm, so the amplification factor is 2.1. The piston front plates exhibit even higher displacement amplitudes in the normal direction: the lower piston's front plate centre moves up and down by 10.8 µm, the upper piston's centre by 14.0 µm. This corresponds to factors of 3.0 and 3.9.

The West-Howlett resonator and the wall pressure ratio

The wall pressure ratio is a crucial property of acoustic resonators for cavitation experiments. Once cavitation happens, it reduces the tension in the surrounding liquid. The creation of cavitation bubbles instantly rises the pressure in their vicinity to

the level of the vapour pressure. Therefore it is a problem if undesired cavitation is induced on structural surfaces inside the resonator. It destroys the desired pressure field without leading to bubble clouds suitable for SF experiments. Low pressure amplitudes on the fluid-structure interface is thus a quality measure for the resonators to be used. In a liquid with a high cavitation strength (which is able to sustain large tensions before rupturing) like acetone, it is the tension limit on the structural surfaces which imposes the limit on how far the maximum pressure amplitude can be pushed up by the resonator. As long as the amplifier and transducer system can still handle an increase in power level, as long as the stability limit inside the liquid bulk is still far away, it is cavitation on surfaces which limits the useful driving power range for experiments based on bubble cluster bursts.[3]

The cavitation strengths of liquid-structure interfaces are material-dependent constants, but the various simulated resonators exhibit many different values of $p_{\max}$ which are the basis for normalising the wall pressure ratio r_{wp}. Why then can these cases be compared by their value of r_{wp}? The comparison is possible because in a real SF experiment a resonator will not be excited with a predetermined voltage amplitude, rather the suitable driving power will be chosen to achieve a desired value of $p_{\max}$, e. g. the neutron-induced cavitation threshold of -7 bar doubled (as indicated in Taleyarkhan's SF experiment protocol, see apdx. D). This means that most resonator setups in this chapter can be directly compared via their wall pressure ratios whereby lower values are better. Only cases where the highest p_{if} occurs on different structural materials are not directly comparable. Unfortunately, the cavitation strengths on the surfaces of glass, steel, or the other materials were not investigated during the experimental resonator characterisation campaigns preceding this work. Furthermore, literature data is sparse because two materials and conditions need to match and unreliable because of the strong dependency on factors like impurities or surface microstructure. These are the reasons why also with respect to r_{wp} the simulation of the West-Howlett resonator is taken as the reference. As described in figures 5.3 and 5.4, the values of r_{wp} are ranging from 25 to 40 %. Therefore, during resonator EAO trials and accompanying experiments wall pressure ratios from 20 to 30 % were taken as the limit of worst acceptable values. The acceptable r_{wp} limit and eventually its material-dependence is one more topic to be re-examined when restarting experimental work in future projects.

In the following sections it will be shown that the fitness evaluation routine offers a lot of flexibility for steering the optimisation process. Through the evaluation routine one can require properties making a resonator suitable for SF experiments with acetone. The EA-optimised resonators show that the r_{wp}-based penalty in the applied setup can effectively ensure that they at least match the West-Howlett resonator model, and that local or global searches with harshened penalty functions are in fact able to yield resonators with much lower wall pressure ratios. If the envisaged sound pressure amplitude is 14 bar in an SF experiment, then wall pressure ratios of 20, 30, and 40 % can be translated into tension states of -2.7, -4.2, and

[3] Therefore it makes sense to put the whole resonator under pressure (positive static pressure offset), an idea pursued by Ross Tessien et al. of Burst Laboratories ®, see chapter 1.3.5 or i. a. [400].

−5.6 bar between liquid and wall.[4]

Reconsidering the prioritisation of $p_{\max}$ and r_{wp}

One can take the above discussion of the $p_{\max}$ and r_{wp} values gained from the FE model of the West-Howlett resonator as an occasion to ask a principal question: How useful is the objective function definition as presented in chapter 4.2 and as applied in most EAO case studies where $p_{\max}$ is the primary goal and the wall pressure ratio-based penalty scheme comes on top? Might it not be better to switch the priorisation? Should the minimisation of r_{wp} perhaps become the primary goal of a resonator optimisation procedure and should $p_{\max}$ be second priority? In follow-up projects it might be interesting to try out alternative goal formulations, other penalisation schemes, acquire capabilities to be able to treat constrained optimisation problems, or to examine pareto fronts. The justifying thoughts behind the approach taken here are the following ones: One could say that the minimisation of r_{wp} represents optimisation under the assumption of unlimited driving power. By contrast, the maximisation of $p_{\max}$ tries to get the strongest resonance peak out of equipment with a given power source and a given level of material damping. The latter approach was deemed to be the more practical one for achieving high-Q resonators with optimised sound field topologies. After all, it was the low-Q signature of the latest experimentally investigated resonator design which rendered it unsuitable for SF experiment repetitions and which furnished the motivation for the subsequent FE modelling and optimisation efforts.

There is yet another way to put it: while the $p_{\max}$ criterion is assumed to be the best driving force to optimise the resonator design when it has to work with the given transducer system, r_{wp} might be a good objective function for a broader optimisation procedure considering also trade-offs in the relative sizing of the transducer versus the resonator (one might be inclined to go to stronger transducer power in a setup where a sufficiently powerful cooling system is provided). It is possible to be happy with a design with very low r_{wp} and and a mediocre $p_{\max}$ performance if one knows that there is a way to achieve desired sound pressure amplitudes by scaling up the driving power and that this way will not encounter problems like limits of cooling and material or assembly stability.

5.3 Geometry A: precision-machined pistons

The first examined new resonator design, depicted in figures 5.5 & 5.6, is referred to as *geometry A*. It is based on the shape of the West-Howlett resonator design which is noticeable by the similarities in the piston shapes. After simple and straightforward design modifications the new geometry avoids many problematic aspects identified during the preceding work at RPI while conserving the shape and the oscillation mode of the liquid volume. The problematic aspects were addressed by the following measures:

[4] assuming a vapour pressure of acetone at 0 °C of 0.093 bar and implying a sound pressure amplitude of 14.2 bar reached upon amplitude doubling after touching the cavitation threshold of −7 bar tension

- No manual glassblowing work is necessary in geometry A because merely a straight section of an industrially manufactured glass tube is used.

- Some piston versions still involve front plates made of glass but the shape of these glass pieces allows their manufacturing by sanding glass blocks.

- All connections are glue connections (epoxy and silicone) instead of assembly work with nuts and bolts.

- The connection between pistons and their equilibration mass, the endplate, is made of a single metal piece (in the West-Howlett design a silicone connection transmits the forces of the piston bouncing against the outer glass hull). The endplates are allowed to become (or can be enforced to stay) massive during the optimisation which offers a way to minimise their displacement.

- Massive endplates with low displacement amplitudes are readily available for drilling outlets and placing nozzles in a way minimising the expected deviations between a 2D-axis-symmetric simulation and future experimental trials.

The parametrised geometry is shown in figure 5.5. Algorithmic optimisation can only be successful if the base geometry and the parametrised design space allow for the necessary vibration degrees of freedom of the structure. Therefore it is important that the piston front plate can be thin or become thick and heavy while the rim can get thin and turn into a soft spring. This is of particular importance because the piston holding tubes cannot move vertically against the rest of the structure. They are fused to the endplates as a measure to prevent unintended tilting of the piston in a real-world setup. On the side of the glass cylinder it is deemed important that for various transducer positions a vibration mode shape can develop which can at the same time beneficially couple to the pressure field in the liquid while exhibiting displacement nodes at the cylinder ends in order to minimise the energy loss through damping in the silicone connections to the endplates.

The subsequent figures present a collection of different optimisation results. As geometry A was the main learning case they also represent different steps of the development process of the EA and its interfacing with the FEM simulation. The samples were chosen for being performant optimised setups achieving a high sound pressure on the one hand, but on the other hand some choices were driven by the will to document important lessons learnt during the process, aspects which should be kept in mind when continuing the SF resonator development in the future. Several of them are deemed to be of more general interest in the context of other optimisation problems in science and engineering. The following list outlines the figures and their associated issues:

- The first couple of pictures represent early EA application trials. Figures 5.7, 5.8, and 5.9 show symmetric versions of the geometry allowing to cut the FE model size in half. Figure 5.7 is the result of a trial where only six parameters were tuned, in figures 5.8 and 5.9 their number was already thirteen. A conflict of interest is being discussed between the goal to maximise the freedom given

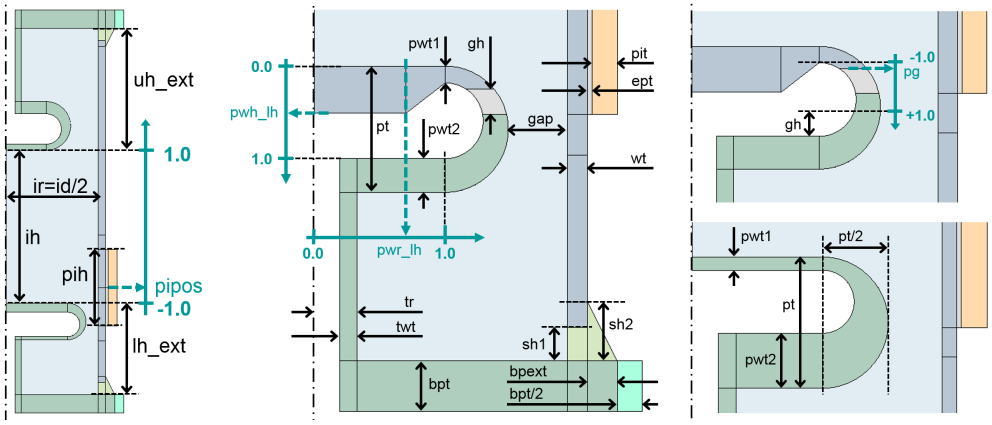


Figure 5.5 Parametrisation of geometry A

The main idea behind this geometry is to conserve the piston shape of the West-Howlett design as part of an assembly of precision-machined metal and glass parts where the pistons cannot tilt. The hollow piston shapes either have to be glued or welded together from parts or manufactured via additive methods. In this design there is no soft connection between a piston and the end plate, they are one piece. The needed vertical displacement of the piston front plates has to arise in conjunction with an internal vibration mode of this solid part. Therefore the parametrisation has to allow for segments of the piston wall to become thin to enable some flexibility for motion patterns where thicker parts serve as frequency-tuning masses. Soft silicone connections exist between the glass cylinder and the end plates. The two close-ups on the right show the two of the three implemented piston variants: (a) a piston front plate of constant thickness made of glass is glued with epoxy to a metal back part versus (not shown), (b) the piston has a glass front plate with two parameters determining a thickness profile, and (c) an all-metal hollow piston.

label	description	value
id	inner diameter of main volume	59.2
ih	inner height of main volume	[50,120]
pit	piezo ring thickness	3
pih	piezo ring height	25
pipos	parameter determining the transducer's vertical position	[-1,-0.5]
ept	epoxy layer thickness	0.5
ext_uh	extension from reflector to base plate (upper half)	[30,60]
ext_lh	extension from reflector to base plate (lower half)	[30,60]
wt	wall thickness of main cylindrical glass wall	2.4
gap	gap between piston rim and inner surface of glass hull	[2,12]
pt	piston thickness	[8,24]
pwt1	piston wall thickness 1 (front/reflector plate)	[1,8]
pwt2	piston wall thickness 2 (back side)	[1,8]
pwr	parameter controlling radius of added weight on reflector plate	[0.1,0.9]
pwh	parameter controlling thickness of added weight on reflector plate	[0.1,0.9]
gt	epoxy glue layer thickness	1
pgp	parameter controlling epoxy glue layer position	0
tr	radius of piston holding tube	3.55
twt	tube wall thickness	1.05
bpt	base plate thickness	6
sh1	height of silicone betw. base plate and glass cylinder	[1,8]
sh2	total height of silicone bead	10
silext	base plate extension supporting silicone bead	[3,15]

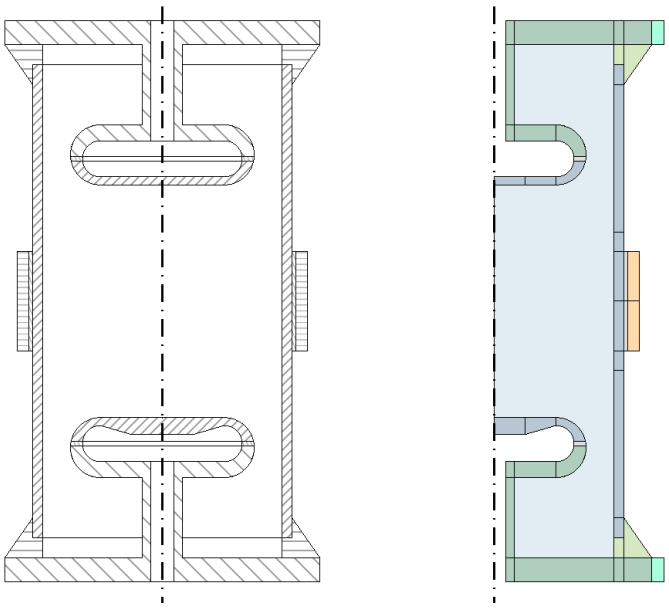


Figure 5.6 Geometry A: full cross section versus simulated geometry
The drawing on the left shows a full cross section of an exemplary instance of resonator geometry A. The geometry on the right is the half cross section upon which the finite element mesh of an axis-symmetric model can be built.

to the optimiser and keeping the condition of comparability between different optimisation results. Furthermore, the penalisation for high pressures on structural walls is motivated.

- Many different mode shapes can be found by applying EAs to the optimisation task. They can differ greatly from the working mode of the West-Howlett resonator. The next three cases show the best alternative vibration modes discovered by EA search. In figure 5.10, it is the simplest case, the fundamental mode of the central liquid volume with one single antinode in the radial and the vertical direction. The sound field is very close to spherical symmetry and the sound pressure on the fluid-structure interfaces is very low. But it has to be noted that the optimiser achieves this through a setup where the radial diameter of the fluid volume is not the same as the vertical diameter; it can be taken as a hint that a wavelength-based resonator layout rule (requiring a match of diameter and height of the central volume with half a wavelength) can only approximate the neighbourhood of optimal setups.

- The case shown in figure 5.11 is a mode with two separate subvolumes predestined for cavitation. Their pressure is oscillating in phase. It is a very simple mode shape but not trivial in the sense that n half wavelengths in the radial direction are superpositioned with m humps along the vertical axis. One can ask the question of how probable the discovery of such a working mode would be when working along conventional patterns without using computer-aided routines of black-box optimisation and in particular random-influenced algorithms going unforseen ways. The opinion that “human design, driven by deep

understanding is better than the lazy approach of black-box treatment" is a justifiable position. But it would be wrong to perceive a conflict between two fundamental approaches and turn this into an argument against considering EA techniques. The above thoughts are presented to support an alternative insight that can be gained from EA application experience: the potential of black-box optimisation techniques to come up with surprising solutions and challenge habitual thought restrictions can be used as an everyday tool to trigger and accelerate deeper problem understanding.

- Figure 5.12 contains an exemplary plot showing the developing fitness distribution achieved by applying THEA with the population merging scheme. By the large fraction of negative scores it illustrates the problematic effect of a linear and unbounded wall pressure-based penalty term.

- Another simple mode shape with two strong bellies, this time out of phase, is depicted in figure 5.13.

- After having presented several alternative mode shapes, figure 5.14 is an example where EA search has discovered the mode shape of the West-Howlett resonator. Both CMA-ES and THEA were able to discover that particular working mode. The discussed case does not represent the trial evaluation routine with mode shape discrimination. The base value for the score calculation is the $p_{\max}$ found anywhere on the vertical axis in the region of interest. This proves that the West-Howlett mode shape can be found without mode shape filtering and that similar sound pressures can be achieved with this shape as compared to the alternative mode shapes. It indicates at the same time that this mode shape is not systematically disadvantaged (and by consequence much harder to find) in the new geometry. It confirms that the parametrised geometry A offers the needed structural degrees of freedom to be able to host the same acoustic pressure field as the West-Howlett resonator.

- The same figure also serves as the exemplary case for explaining an implementation flaw of the score evaluation routine used in most cases for geometries A and B.

- That the two applied EAs of CMA-ES and the in-house algorithm THEA were both able to find some similar solutions is shown in figure 5.15.

- Of course there are also drawbacks when using a black-box optimisation tool. An important one is the effort necessary to make the subroutine executing one evaluation call quite robust. Occurring errors act as drag on the optimisation efficiency for sure. But figure 5.16 shows something worse: an example of the optimiser using errors or model artefacts to maximise the fitness function but not solving the engineering problem. It is the final result of a case where chromosomes representing models with errors took over the gene pool and made the rest of the optimisation run useless. The case shows that the user of black-box optimisation routines sometimes has to take special care to make sure that the goal setting as well as the boundary conditions for the optimiser

are in line with the engineering problem. At the same time there is a positive message: the case illustrates the global search efficiency of the employed EA, i. e. the ability to identify small regions holding exceptionally promising design variants in a vast and structure-rich design space.

- The downhill-simplex search in its original form ignores search domain boundaries. They have to be introduced artificially. Figure 5.17 shows a comparison of unbounded versus bounded downhill-simplex search originating from the same starting point but leading to different results.

- Figure 5.18 adds the parameter histories to the discussion of the two downhill-simplex runs of figure 5.17. They illustrate the strong coupling between the input parameters in their influence on the objective function.

- Two local search results with two different starting points are shown in figures 5.17 & 5.19. The starting points are the best chromosome of an evolved EA population and the best one of the random set with which this particular evolution run started. The fact that these two local search runs do not end up on the same fitness function hill is just another proof of existing barriers between local optima and the necessity to treat the resonator optimisation problem with a global search algorithm. It says that a global Monte-Carlo scan with subsequent local search is not sufficient, it is an inferior approach compared to EA optimisation for the given challenge.

- The geometry setup in figure 5.21 (case 212) together with the one shown in figure 5.17 (case 138) show results gained with the latest EA setup and the score evaluation routine involving mode shape discrimination. The fitness function was calculated as described in figure 4.7 which means that only out-of-phase pressure antinodes appearing above the upper rim of the transducer were taken into account and multiplied with a sine window. The windowing penalises all those mode shapes for which the peak pressure does not occur near the middle of the region of interest, i. e. between the upper transducer rim and the upper piston's front plate. This is deemed to lead to a substantial enlargement of the West-Howlett mode shape's attractor region in the search space, and indeed, only such mode shapes came forth as winners of the evolutionary competitions in these cases. Case 212 represents an additional EAO run performed after correcting the flawed computation of the wall pressure ratio and the penalty. The effect is clearly visible in the pressure contour plots and the lowered value of the wall pressure ratio. The FEM simulation results of cases 138 and 212 show that when geometry A is used to reproduce the working mode of the West-Howlett resonator, it is able to develop the advantageous feature of a wall displacement amplification within the cylindrical glass wall.

These figures[5] and their discussions fill the following pages. A summary will follow at the end of the chapter discussing all examined geometries.

[5]The figures are often labelled with case numbers. These case numbers are unnecessary information to the general reader, but they may give better orientation to successors on the project who will inherit the stored data.

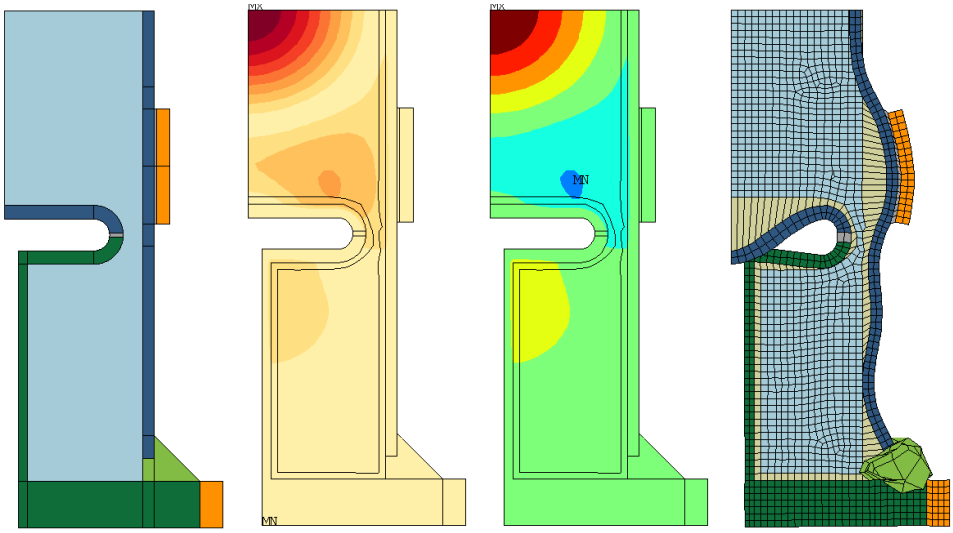


Figure 5.7 Geometry A: case 73
For the symmetric version of the geometry only half the model needs to be simulated. In this early EA trial only six parameters were tuned. The in-house EA, THEA, was used for optimisation: four populations of chromosomes were random-initialised, evolved over six generations, then merged. THEA was continued with the merged population for 40 generations, whereby the goal was to simply maximise $p_{\max}$. THEA in this early setup was lacking the DE tier (differential evolution) and used a mutation operator with $P = 0.6$. The population size was $N_{\text{population}} = 80$. This means the total cost of the EA optimisation was 5120 evaluation calls as can also be seen from the table with the optimisation key facts below. From the resonator specs table it can be seen that the resonator achieves a sound pressure of 37.89 bar. This might seem high in comparison to the optimised West-Howlett resonator models shown in figures 5.3 and 5.4 but it has to be remembered that in the symmetric model one single transducer turns into two if it shifts away from the centre position. It can also be noted that the sound field and vibration mode of the above example are the symmetric version of the mode shape of the West-Howlett resonator: the transducer is close to the piston, there is one weak in-phase antinode in front of the piston and the envisaged cavitation region is part of an out-of-phase antinode above it.

(list of parameters: *ih*, *pipos*, *gap*, *pt*, *pwt1*, *pwt2*)

resonator specs	
f_{res}	21 680 Hz
$p_{\max}$	37.89 bar
$p_{\text{upc}}, p_{\text{lpc}}, p_{\text{p}}$	2.01, 2.01, 12.45 bar
p_{wall}	6.42 bar
r_{wp}	0.33
Q_{mech}	254

optimisation key facts	
EA	THEA
scanned	18.5-22.5 kHz
f_{obj}	$p_{\max}$
$N_{\text{population}}$	80
generations	$4 \cdot 6 + 40 = 64$
N_{eval}	5120
local search	–
N_{eval}	–

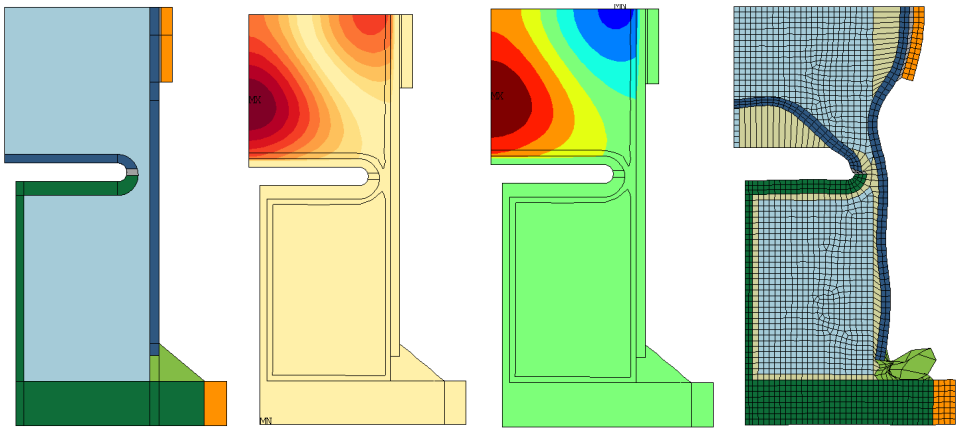


Figure 5.8 Geometry A: case 74a

This and the next figure show two more examples of solutions found by simple and short evolution runs, simple because the goal was solely the maximisation of $p_{\max}$ without penalty, short because the random-initialised populations evolved merely over five generations. Together with figure 5.7 they illustrate that a variety of different mode shapes can be found where the pressure peak is internal as desired, which do not directly fall into the highly undesirable category with the pressure peak attaching closely to a solid part (as in figure 4.3), and which thus become interesting as SF resonator working modes if further optimised. The solution depicted above is unusual, on the one hand because there are two amplitude antinodes along the central axis which oscillate in phase, on the other hand because of the high achieved sound pressure of 60.90 bar. However, severe drawbacks of this simple EAO setup are revealed as well. There is a problem arising from gearing the optimiser solely at maximising the centreline pressure: in the the pressure field snapshot above, the dark blue colour almost touches the glass wall and the red colour almost touches the piston (the second-last colours in blue and red correspond to $r_{\mathrm{wp}} > 0.55$). At sound pressure amplitudes up to 15 bar as required for the SF experiment this means strong tension in the liquid on the glass wall behind the transducer and in front of the pistons where it is not desired. It translates into a wall pressure ratio of 0.68. Such results and worse EAO outcomes not shown here have revealed the need for more sophistication with the ranking of EA population members, so that setups with high wall pressure ratios become less likely to win the evolutionary contest.

(list of tuning parameters: *ir*, *ih*, *pipos*, *gap*, *pt*, *ext*, *wt*, *pwt1*, *pwt2*, *bpt*, *gh*, *sh1*, *silext*)

resonator specs	
f_{res}	21 517 Hz
$p_{\max}$	60.90 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	41.69, 41.69, 41.69 bar
p_{wall}	33.04 bar
r_{wp}	0.68
Q_{mech}	1091

optimisation key facts	
EA	THEA
scanned	18.5-22.5 kHz
f_{obj}	$p_{\max}$
$N_{\mathrm{population}}$	80
generations	6
N_{eval}	480
local search	–
N_{eval}	–

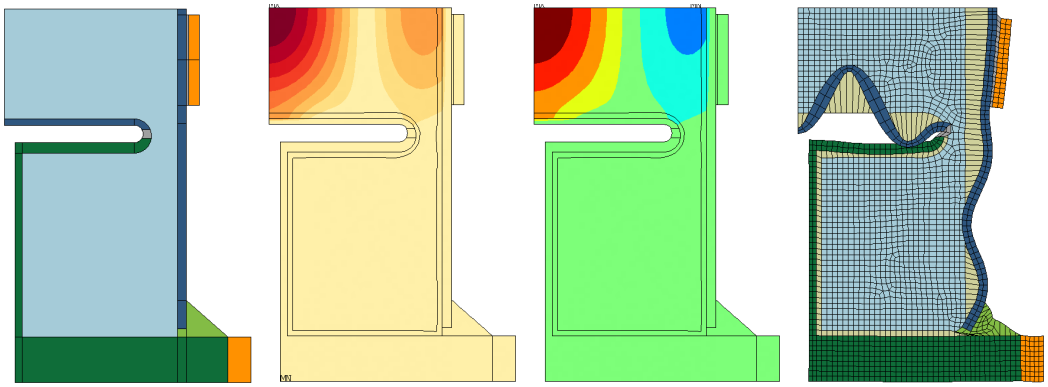


Figure 5.9 Geometry A: case 74b

Depicted is a third example of a high quality solution from an EA population optimising the symmetric model. It is again a solution of very high $p_{\max}$. The characteristics of this resonator setup are a very short inner height, allowing a pair of closely packed transducers covering almost the entire distance from piston to piston. Comparing such a case with the single-transducer setups is problematic. There is a conflict between maximising the degrees of freedom given to the optimisation routine and maintaining comparability.

In this case there are two tuning parameters influencing the sound pressure performance directly through the transducer size. The first is the already mentioned parameter *pipos*. When it is set to zero then the transducer is placed in the central position and only half of it is represented in the symmetric FE model. Moving the transducer downwards ($pipos < 0$) first makes it grow in size (like in fig. 5.8) and finally splits it into two separate piezo rings (as in fig. 5.7 or here). The other parameter is *ir*, the inner radius of the glass cylinder, which directly influences the transducer radius. Enlarging *ir* means increasing also the available mass of piezoelectric ceramic exciting the resonator. This is a problem for the comparability of different resonator design solutions. It is not clear any more whether a high $p_{\max}$ reflects a beneficial sound field topology and distribution of stored elastic energy across the various structural parts or simply the impact of a larger transducer. For that reason it was decided to keep the transducer size fixed in later optimisation runs.

(list of tuning parameters: *ir*, *ih*, *pipos*, *gap*, *pt*, *ext*, *wt*, *pwt1*, *pwt2*, *bpt*, *gh*, *sh1*, *silext*)

resonator specs	
f_{res}	18 873 Hz
$p_{\max}$	58.14 bar
$p_{\text{upc}}, p_{\text{lpc}}, p_{\text{p}}$	25.31, 25.31, 25.31 bar
p_{wall}	20.89 bar
r_{wp}	0.44
Q_{mech}	885

optimisation key facts	
EA	THEA
scanned	18.5-22.5 kHz
f_{obj}	$p_{\max}$
$N_{\text{population}}$	80
generations	6
N_{eval}	480
local search	–
N_{eval}	–

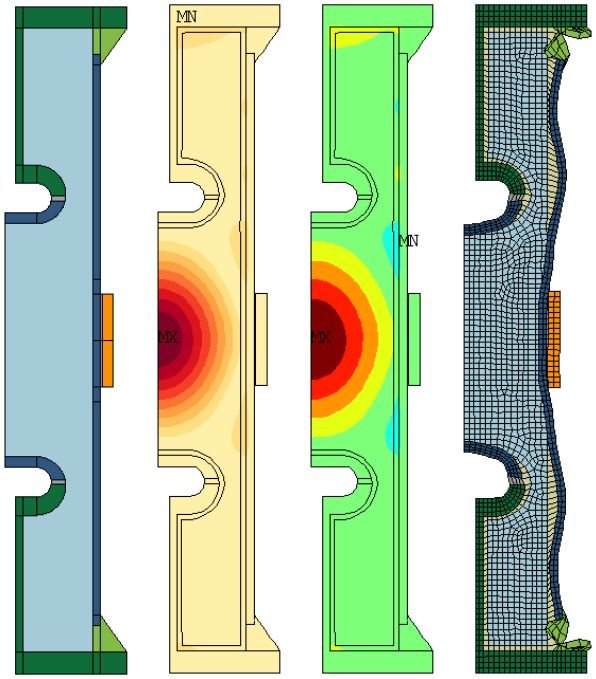


Figure 5.10 Geometry A: case 95a

Several optimisation runs were conducted with the same set of tuning parameters as discussed in the last figure but with the whole FE model (no horizontal mirror plane), where the resonator geometry is basically symmetric and only the piezo ring position is allowed to break the symmetry. This led to several cases where the sound field adopted a symmetric shape during the optimisation. Two of these cases are shown in this and the following picture. Here, a unimodal sound pressure field is visible. The achieved sound pressure of 25 bar is above the reference case of the manually optimised Wet-Howlett design and below the algorithmically tuned version. The notable feature of the pressure field is its roundness. There is a mismatch between the inner diameter of 48 mm and the inner height with 63 mm. By consequence, the sound pressure isocontours are of elliptical shape, although with less unequal radii than the diameters of the liquid volume itself. The piston front plates are quite sturdy. They move inwards in phase with the piezo ring and with a similar amplitude as can be seen from the displacement plot. The vibration mode of this setup is accompanied by a relatively low wall pressure ratio of 20 %.

(list of parameters: *ir*, *ih*, *pipos*, *gap*, *pt*, *ext*, *wt*, *pwt1*, *pwt2*, *gh*, *sh1*, *silext*)

resonator specs	
f_{res}	22 644 Hz
$p_{\max}$	24.96 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	2.00, 0.93, 3.42 bar
p_{wall}	5.00 bar
r_{wp}	0.2
Q_{mech}	253

optimisation key facts	
EA	THEA
scanned	18.5-22.5 kHz
f_{obj}	$p_{\max}$ with $\tilde{f}_{\mathrm{penB}}(p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{wall}})$
$N_{\mathrm{population}}$	80
generations	$4 \cdot 14 + 84 = 140$
N_{eval}	11 200
local search	–
N_{eval}	–

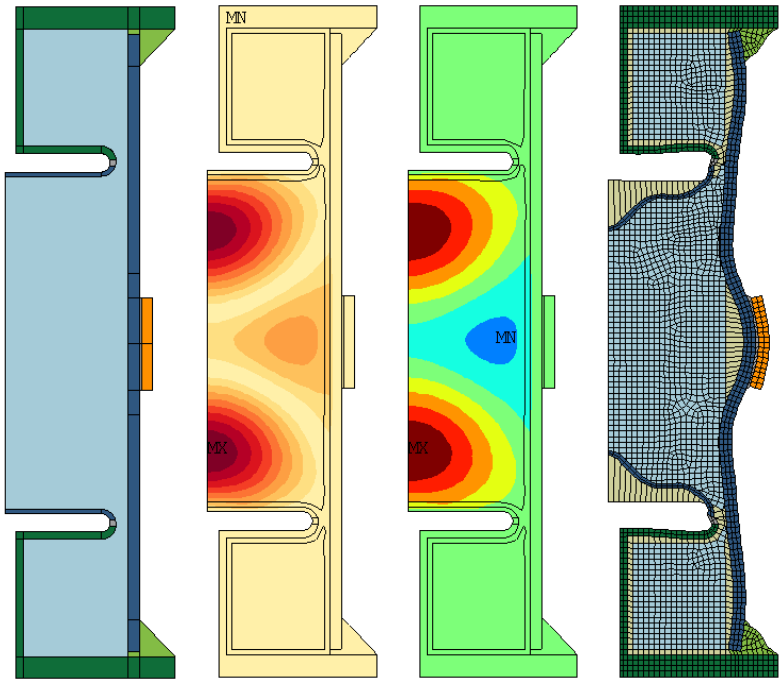


Figure 5.11 Geometry A: case 95b

This case can be seen as the inversion of case 73 depicted in figure 5.7. Again there is a weakly blue region on the central axis in the pressure snapshot and a strongly red region next to it. Only this time the transducer and the weak in-phase antinode are in the centre whereas the out-of-phase pressure amplitude regions are by the side, i. e. next to the pistons. Even the triangular shape of the cyan area is similar. The pistons became very wide and flat. Their glass front plates have been tuned very thin by the EA. With 1.1 mm this parameter has been pushed almost to the limit of the allowed range. This resonator exhibits a $p_{\max}$ of more than 32 bar and is in this respect the most performant single-transducer setup. Its wall pressure ratio of 27 % is similar to the West-Howlett resonator model and thus at the limit of the declared acceptable range. A score history plot corresponding to this particular optimisation run is shown in figure 5.12. In case 95 there are two outcomes labelled (a) and (b) because the proto-populations were merged a second time entailing a different subsequent evolution path of the final population.

(list of parameters: *ir*, *ih*, *pipos*, *gap*, *pt*, *ext*, *wt*, *pwt1*, *pwt2*, *gh*, *sh1*, *silext*)

resonator specs	
f_{res}	22 192 Hz
$p_{\max}$	32.30 bar
$p_{\text{upc}}, p_{\text{lpc}}, p_{\text{p}}$	4.91, -?-, 8.69 bar
p_{wall}	8.38 bar
r_{wp}	0.27
Q_{mech}	674

optimisation key facts	
EA	THEA
scanned	18.5-22.5 kHz
f_{obj}	$p_{\max}$ with $\tilde{f}_{\text{penB}}(p_{\text{upc}}, p_{\text{lpc}}, p_{\text{wall}})$
$N_{\text{population}}$	80
generations	$4 \cdot 14 + 63 = 119$
N_{eval}	9520
local search	–
N_{eval}	–

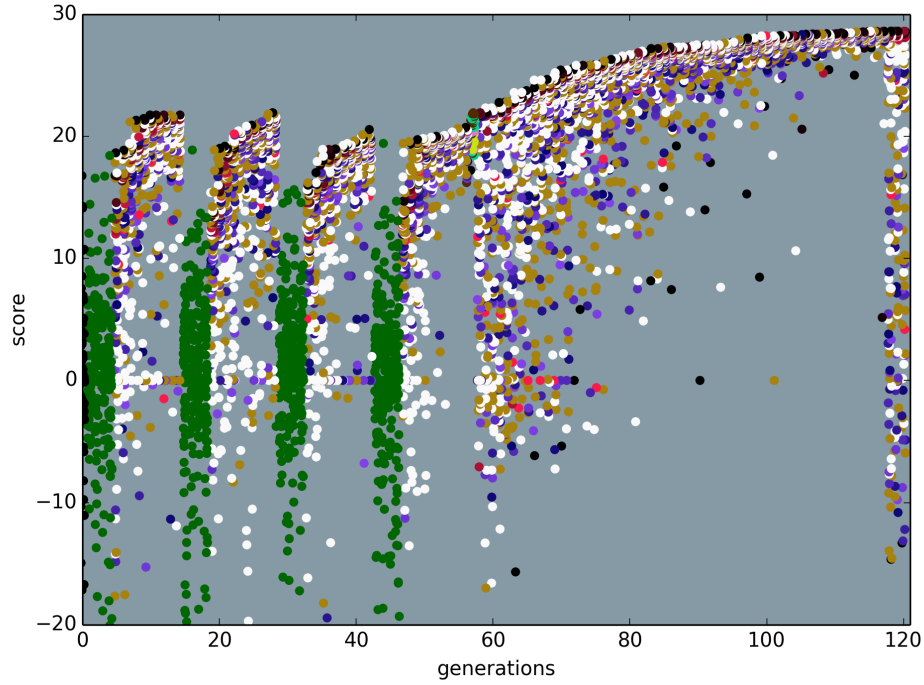


Figure 5.12 Geometry A: case 95b score history
This scatter plot shows the scores occurring in the population versus time. The green dots represent random trials and their occurrence shows the initialisation phases of the four proto-populations. Each proto-population, after having been assembled by the combination of the best random trials, evolves for 10 generations. The scores of the merged final population occur in generation 58. The effect of the decaying mutation step size parameter (cooling rate $\gamma = 0.04$) can be observed from that point onwards. This particular score history plot differs from all the others in one aspect: three generations before stopping the EA the mutation step size parameter has been manually reset to a large value for checking whether the population would jump into a different local optimum. This did not happen and the top score stayed constant.

Another feature visible in this plot is that the wall pressure-based penalty in the form of an additive uncapped term in the objective function (eq. 4.1) creates a lot of trials with negative scores in this case. That highly penalised trials with a strong sound pressure performance are ranked behind the trials with the weakest sound pressure amplitudes (both weakly and strongly penalised ones) is surely a drag on the optimisation efficiency. It was a reason for further changes to the penalty formulation.

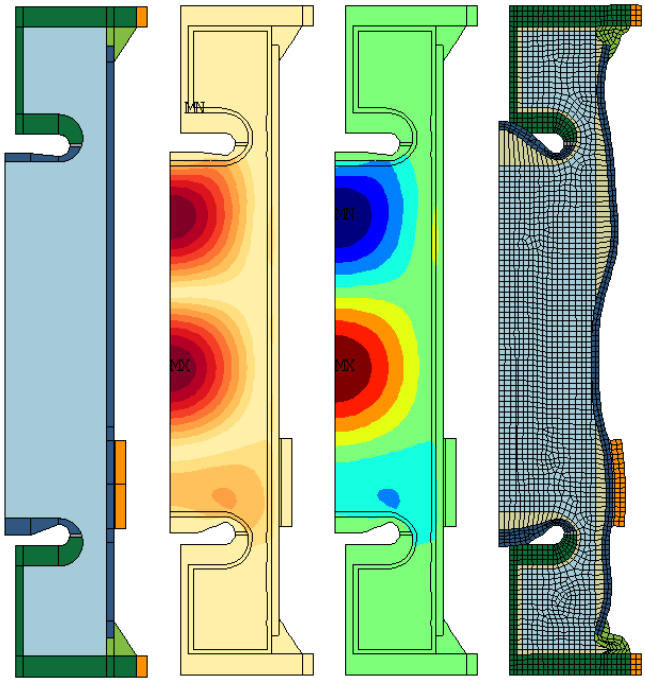


Figure 5.13 Geometry A: case 109a

In this setup two strong antinodes suitable as cavitation regions oscillate out of phase. A similarity to the cases described before lies in the fact that the transducer is exciting on the height of a weak antinode neighbouring the strong antinodes on the vertical axis. In both geometries of case 95 (figures 5.10 & 5.11) the epoxy layer has to serve as a spring for the bouncing and deforming piston front plates. Here the front plate geometry has been made a little bit more complex which also increases the number of tuning parameters. Now there are two thicknesses and a transition region implemented allowing a thick central part to be surrounded by a ring region of thin glass allowing easier bending. It creates some design freedom to lower the degree of deformation and energy storage burdening the epoxy layer. Concerning the possible degree of asymmetry between upper and lower piston, many design parameters are still kept the same. As the suffixes "uh" and "lh" in the parameter list indicate, only the thickness of the central region and the length of the piston holding tube can adopt different values in the upper and lower half. The piston diameter (*gap*), the back wall thickness (*pwt2*), the thin front wall thickness (*pwt1*), all these parameters are the same for both pistons. It can be assumed that for the optimiser there are still limitations when trying to tune different displacements and different acoustic impedances at one common resonance frequency.

(list of parameters: *ir*, *ih*, *pipos*, *gap*, *pt*, *ext_uh*, *ext_lh*, *wt*, *pwt1*, *pwt2*, *gh*, *sh1*, *silext*, *pwh_uh*, *pwh_lh*, *pwr*)

resonator specs	
f_{res}	22 176 Hz
p_{max}	18.77 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	2.86, 1.70, 6.89 bar
p_{wall}	3.22 bar
r_{wp}	0.37
Q_{mech}	480

optimisation key facts	
EA	(40,80)-CMA-ES
scanned	18.5-22.5 kHz
f_{obj}	p_{max} with $\tilde{f}_{\mathrm{penB}}(p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{wall}})$
$N_{\mathrm{population}}$	80
generations	43
N_{eval}	3440
local search	–
N_{eval}	–

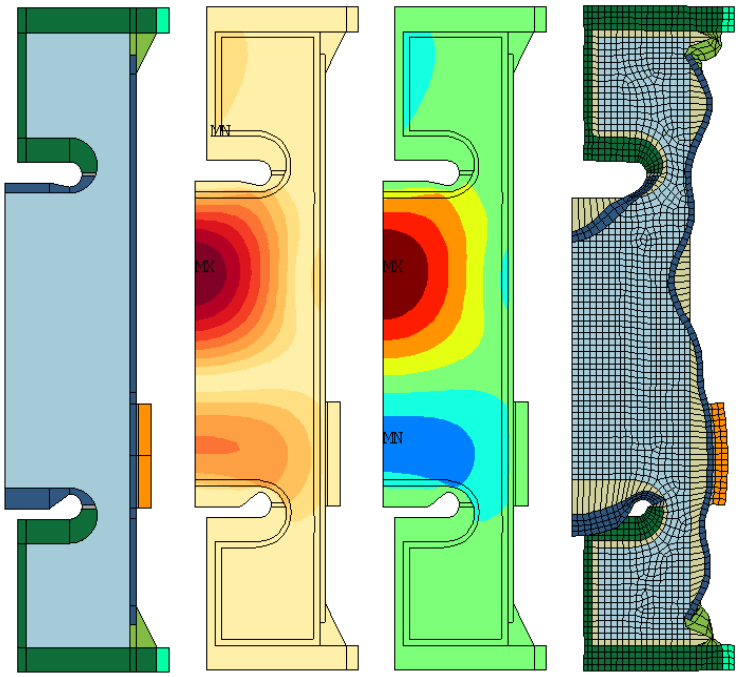


Figure 5.14 Geometry A: case 109b

Figures 5.7 - 5.13 show several different local performance optima of geometry A found by EAO and exhibiting mode shapes appearing suitable for acoustic cavitation. The setup depicted here presents a case where an EAO resulted in the very same working mode and equivalent pressure field as in the West-Howlett resonator design. It shows that this working mode can also exist in the new design, that it is associated with a good $p_{\max}$, and that it can be found by EA tuning with a relatively simple fitness function implementation without mode shape discrimination. In fact, both CMA-ES and THEA were able to find this mode with the simple fitness function relying on the direct $p_{\max}$ found in the region of interest and a wall pressure penalty according to equation 4.1. The result of an optimum search with CMA-ES is shown here.

Moreover, the above plots are suitable for the necessary explanation about a shortcoming of the fitness function setup used in most cases of tuning geometry A. The fact that the central regions of the front plates have large vertical displacement amplitudes in comparison to the outer rims has an influence on the pressure fields visible very well in the pressure amplitude and snapshot plots. It has the consequence of making the colour contour lines close to the pistons concave. Note that the sound pressure amplitude on the surface of the front plates is larger near the rim and lower in the centre. For a proper calculation of the penalty all parts of the fluid-structure interface should be taken into account. Unfortunately, in the optimisation runs of geometry A this was never really the case. Geometry A was the learning problem, and due to unawareness the penalty function was $\tilde{f}_{\text{pen}}(\max(p_{\text{upc}}, p_{\text{lpc}}, p_{\text{wall}}))$ most of the time – and involved only the piston front centre points – instead of the correct formulation $\tilde{f}_{\text{pen}}(p_{\text{if}}) = \tilde{f}_{\text{pen}}(\max(p_{\text{p}}, p_{\text{wall}}))$. The latter involves the peak pressure amplitude found anywhere across the entirety of all piston surfaces, denoted as p_{p}. During this EAO run, the depicted design was thus evaluated based on $r_{\text{wp}} = p_{\text{wall}}/p_{\max} = 0.21$ instead of $r_{\text{wp}} = p_{\text{p}}/p_{\max} = 0.33$. As said before, the concave pressure contours are a consequence of the piston motion. But under the condition of the applied score evaluation, the strong expression of the feature can also be understood as a way for the optimiser to achieve the maximisation of the objective function.
(list of parameters: *ir*, *ih*, *pipos*, *gap*, *pt*, *ext_uh*, *ext_lh*, *wt*, *pwt1*, *pwt2*, *gh*, *sh1*, *silext*, *pwh_uh*, *pwh_lh*, *pwr*)

resonator specs	
f_{res}	20 286 Hz
$p_{\max}$	25.87 bar
$p_{\text{upc}}, p_{\text{lpc}}, p_{\text{p}}$	4.20, 4.63, 8.62 bar
p_{wall}	5.37 bar
r_{wp}	0.33
Q_{mech}	437

optimisation key facts	
EA	(40,80)-CMA-ES
scanned	18.5-22.5 kHz
f_{obj}	$p_{\max}$ with $\tilde{f}_{\text{penB}}(\max(p_{\text{upc}}, p_{\text{lpc}}, p_{\text{wall}}))$
$N_{\text{population}}$	80
generations	102
N_{eval}	8160
local search	downhill-simplex
iterations	350

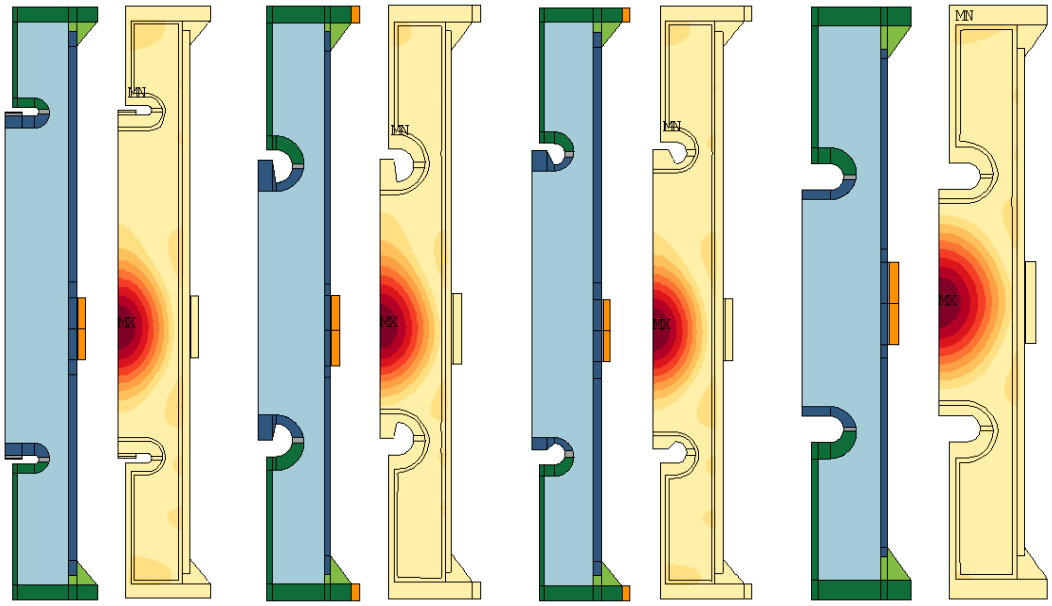


Figure 5.15 Solutions found by CMA-ES and THEA
Various variations of symmetric resonator setups with one single pressure antinode and the transducer in the middle have been found by both applied EAs, CMA-ES and THEA. Four cases are depicted which show that both EAs found local optima with sound pressures between 20 and 25 bar. The four samples also represent trials with different piston geometry variants: the first one with weight plates glued onto the back sides of piston front plates of glass, with thickened front plates in the next two cases, and with the simplest geometry of flat front plates on the right.

resonator specs	
f_{res} (1st setup)	21 433 Hz
p_{max} (1st setup)	21.34 bar
f_{res} (2nd setup)	22 954 Hz
p_{max} (2nd setup)	24.25 bar
f_{res} (3rd setup)	22 151 Hz
p_{max} (3rd setup)	21.26 bar
f_{res} (4th setup)	22 644 Hz
p_{max} (4th setup)	24.96 bar

optimisation key facts	
f_{obj}	p_{max} with $\tilde{f}_{\text{penB}}(p_{\text{upc}}, p_{\text{lpc}}, p_{\text{wall}})$
$N_{\text{population}}$	80
EA (1st setup)	(40,80)-CMA-ES
N_{eval} (1st setup)	5120
EA (2nd setup)	(40,80)-CMA-ES
N_{eval} (2nd setup)	9600
EA (3rd setup)	THEA
N_{eval} (3rd setup)	8320
EA (4th setup)	THEA
N_{eval} (4th setup)	11 200

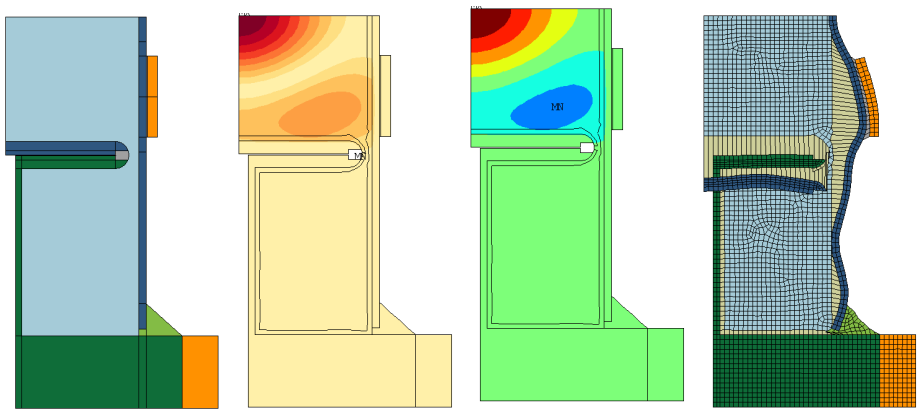


Figure 5.16 Winning the competition with unphysical solutions
A design is plotted where the optimisation algorithm achieved a sound pressure of $p_{\max} = 100\,\text{bar}$ by means of exploiting an error in the FE modelling routines. The error has the effect that the epoxy layer fixing the piston front plate is not meshed. There is no solid connection between the front plate and the rest of the piston. This makes it possible for the front plate to exhibit a large vertical displacement amplitude with little deformation and damping. This way, the meshing error is being exploited for a drastic increase of the sound pressure performance. As a solution to the original engineering problem the model is unphysical and useless. The optimiser was nevertheless tuning a functioning FEM simulation. In a way it has outsmarted the task definition.

Such modelling errors were handled in subsequent optimisation runs of the same geometry by the implementation of additional security checks resulting in a negative score upon detection of unmeshed areas. While this blocks the encountered way in which unphysical solutions can win the evolutionary competition, it does not resolve the issues (a) of drag on the optimisation process inflicted by wasteful trial evaluations ending with error codes and (b) of an augmentation of the search task difficulty due to error-generating regions acting as barriers in the search space. This is the reason why the robustness of the parametrised simulation should be treated as a primary goal from the very beginning.

However, the emergence of solutions with such artefacts can at the same time be seen in a positive light, as a proof for the efficiency of the evolutionary algorithm in exploring the possibilities offered by the design space and in detecting and improving all kinds of available mechanisms allowing the maximisation of the solution fitness.

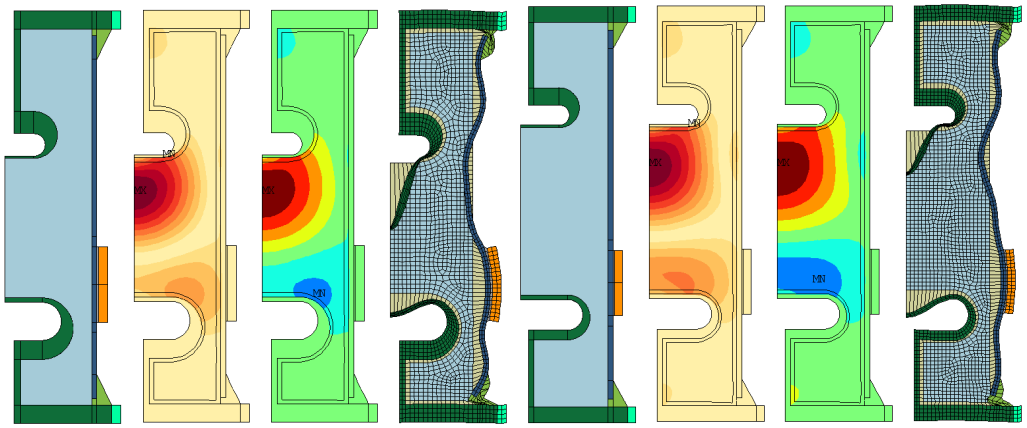


Figure 5.17 Geometry A: case 138, unbounded vs bounded local search
The downhill-simplex algorithm normally does not respect search domain bounds. However, it is easy to penalise out-of-bounds trials artificially. The two geometries above are the results of two downhill-simplex optimisation runs, one without bounds (left) and one with bounds enforcement (right). Their common starting point was a geometry setup found by EA search with a $p_{\max}$ of 22.7 bar. The achieved pressure increases, only 0.3 bar in the bounded search and over 8 bar in the free run, reveal the drastic difference in outcome. Where does this difference come from, and what does it tell?
One explanation attempt can be made based on reminding figure 5.11 where very thin piston front plates oscillate with a large displacement amplitude. Lower bounds are often intended as minimal material thicknesses. Not respecting these limits means unrealistically thin structural parts can appear which would break in reality, but at the same time it is a way to reduce the energy stored in deformed solid material. Everything that is not stored in the structure cannot be lost in the structure, so it is also a way to reduce energy losses through damping.
This kind of outsmarting of the resonator design optimisation task with physically valid but unpractical solutions is deemed a valid explanation approach because it has been observed in other cases. However, in the current case it is not a sufficient explanation. The two thickness parameters shrinking below their lower bound are the thickness of the outer glass wall *wt* and the thickness of the upper piston's front plate *pwt1_uh*. They do not decrease below 95 % of the lower bound in the converged local search, and such little change cannot explain the drastic increase of $p_{\max}$. The proper explanation can be found when examining a different kind of diagram, the change of the tuning parameters x_i during the local search shown in figure 5.18. The free-running simplex search enjoys a freedom the EA did not have and some parameters leave the formerly bounded domain to explore new terrain. This triggers a transition towards a wholly different design point. All tuning parameters have to shift in the course of the transition. It is a useful indication of how strongly coupled the design parameters are.
As the optimisation key facts table shows, this is the first presented result of an EAO run with strong mode shape discrimination. Minimising $\mathrm{Im}(p)$ means looking only for out-of-phase antinodes.
(list of parameters: *ir*, *ih*, *pipos*, *gap_uh*, *gap_lh*, *pt_uh*, *pt_lh*, *ext_uh*, *ext_lh*, *wt*, *pwt1_uh*, *pwt1_lh*, *pwt2_uh*, *pwt2_lh*, *sh1*, *silext*)

resonator specs (unbounded LS, left)	
f_{res}	22 717 Hz
$p_{\max}$	31.38 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	7.15, 4.65, 20.43 bar
p_{wall}	7.26 bar
r_{wp}	0.65
Q_{mech}	466

optimisation key facts (unbounded LS, left)	
EA	THEA
scanned	18-24 kHz
f_{obj}	$max(-Im(p) \cdot \text{sine window})$ with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{wall}})$
$N_{\mathrm{population}}$	80
generations	40
N_{eval}	3200
local search	downhill-simplex
iterations	626

resonator specs (bounded LS, right)	
f_{res}	18 821 Hz
$p_{\max}$	23.01 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	5.70, 5.47, 14.19 bar
p_{wall}	5.57 bar
r_{wp}	0.62
Q_{mech}	423

optimisation key facts (bounded LS, right)	
EA	THEA
scanned	18-24 kHz
f_{obj}	$max(-Im(p) \cdot \text{sine window})$ with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{wall}})$
$N_{\mathrm{population}}$	80
generations	40
N_{eval}	3200
local search	downhill-simplex
iterations	405

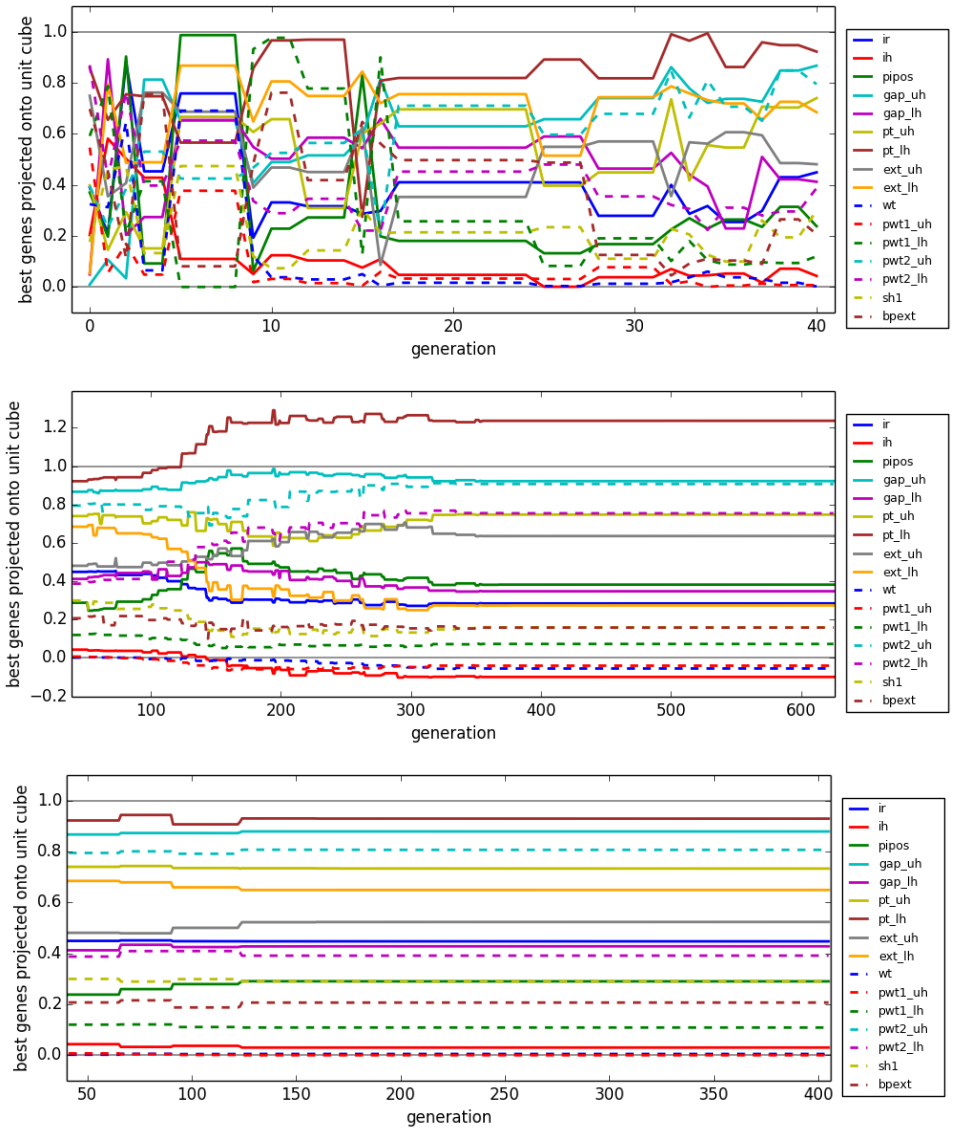


Figure 5.18 Geometry A: case 138, development of best chromosome
The above plots show the development of the best chromosome in the population from iteration to iteration of the optimisation algorithms. For better interpretability the parameters $x_i \in [a_i, b_i]$ have been all mapped onto the interval $[0, 1]$. The first plot shows the 40 generations of the initial EA optimisation. The second plot shows a subsequent unbounded downhill-simplex optimisation. It can be seen that four of the sixteen parameters leave the search domain. The third plot shows another downhill-simplex optimisation, but this time seeing bad scores whenever stepping beyond bounds. In that run the local search modifies the chromosome just a little, and also in terms of $p_{\max}$ it achieves only a small improvement from 22.73 to 23.01 bar (see fig. 5.17). It is interesting to see that of the four parameters going beyond bounds in the free local search only two hit the walls in the bound search. A close look back at the middle plot reveals that the other two parameters hit the limits only after the first two have already left the domain. This is another indicator showing how tightly the resonator's design parameters are correlated. In fact, every single parameter is involved in the transition from one to another resonator layout triggered by the slight search bound transgression of just two parameters. It clearly justifies the classification of the resonator design problem as a highly nonseparable problem.

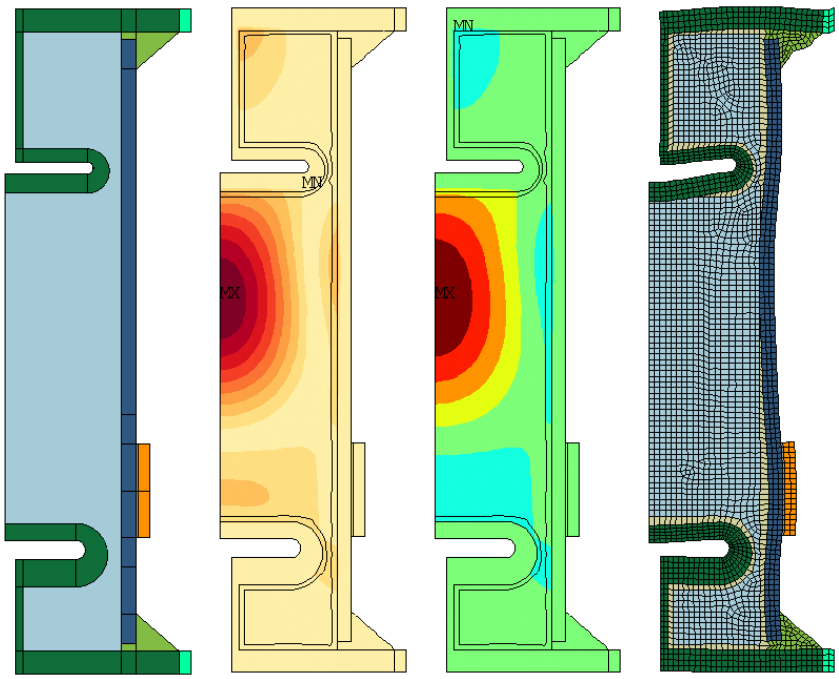


Figure 5.19 Geometry A: case 138, pure local search

Case 138 is the first one in the row of discussed optimisation runs where the fitness evaluation scheme involving mode shape discrimination described in figure 4.7 was applied. Under these conditions no other mode shapes than the intended one were found in EA searches. That scheme looks for a high pressure signal amplitude only above the transducer and below the upper piston. It penalises modes with the pressure peak lying far from the middle of that region of interest and by scanning the imaginary part of the pressure signal and looking for the strongest negative peak it considers only out-of-phase antinodes. This effectively gears the optimiser to tune the resonator in such a way as to adopt the mode shape observed in the West-Howlett resonator. One might say that this expands the attractor region of that mode in the search space. If there is one fitness mountain in the search space which dominates the landscape, so that EA searches do not converge any more on mediocre or in other ways different solutions, then one can ask whether the expensive EA search is still necessary at all, whether a local search algorithm would not suffice for addressing the simple hill-climbing task. Such a check was made by conducting one optimisation run solely with the downhill-simplex algorithm. That search was started with the best chromosome from the same random population with which the EA had been initialised. The simplex was initialised with an edge length of 0.05 times the search domain width in the respective direction. During the simplex run the peak sound pressure increased from 10.0 to 10.1 bar while the score increased by 19 % due to a lowering of the penalty. In the course of centering on a local optimum p_{upc} and p_{wall} became perfectly balanced. Search domain boundaries were not enforced in this case. By the weak sound pressure performance in comparison to both optimisation results in figure 5.17 it is clearly revealed that this type of local search is not suitable for tackling the parameter tuning problem.

(list of parameters: *ir*, *ih*, *pipos*, *gap_uh*, *gap_lh*, *pt_uh*, *pt_lh*, *ext_uh*, *ext_lh*, *wt*, *pwt1_uh*, *pwt1_lh*, *pwt2_uh*, *pwt2_lh*, *sh1*, *silext*)

resonator specs	
f_{res}	19 822 Hz
p_{max}	10.08 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	3.04, 1.75, 3.21 bar
p_{wall}	3.04 bar
r_{wp}	0.32
Q_{mech}	301

optimisation key facts	
scanned	8.8-9.9 kHz
f_{obj}	$max(-Im(p) \cdot \text{sine window})$ with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{wall}})$
local search	downhill-simplex
iterations	581

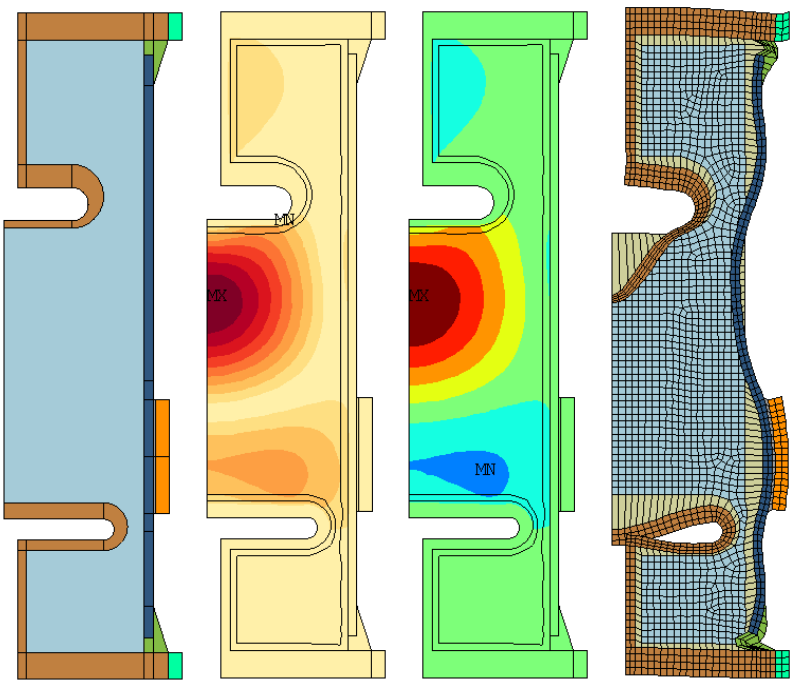


Figure 5.20 Geometry A: case 140

The above setup is the result of tuning the single-piece metal piston variant of geometry A with THEA and downhill-simplex search. Aluminium is used as the piston material whereas in case 138 it was steel. With 29.3 bar a similar sound pressure can be achieved. The local search was conducted with search domain boundaries enforced in order to avoid unpractically thin piston front plates. This setup represents another result based on the flawed penalty computation ignoring the piston rims. It is shown here as a baseline for the comparison with the case 212 presented in the next figure. Case 212 represents a later added EAO run with a corrected wall pressure ratio measurement and penalty calculation. In the setup shown here, the flawed routine allowed the development of elevated interface pressures around the ignored piston rims which goes in conjunction with markedly concave pressure isocontours facing the central regions of the piston front plates.

(list of parameters: *ir*, *ih*, *pipos*, *gap_uh*, *gap_lh*, *pt_uh*, *pt_lh*, *ext_uh*, *ext_lh*, *wt*, *pwt1_uh*, *pwt1_lh*, *pwt2_uh*, *pwt2_lh*, *sh1*, *silext*)

resonator specs	
f_{res}	21 142 Hz
p_{max}	29.26 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	3.32, 4.38, 10.90 bar
p_{wall}	5.57 bar
r_{wp}	0.37
Q_{mech}	497

optimisation key facts	
scanned	18-24 kHz
EA	THEA
f_{obj}	$max(-Im(p) \cdot \text{sine window})$ with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{wall}})$
$N_{\mathrm{population}}$	80
generations	$4 \cdot 14 + 24 = 80$
N_{eval}	6400
local search	downhill-simplex
iterations	344

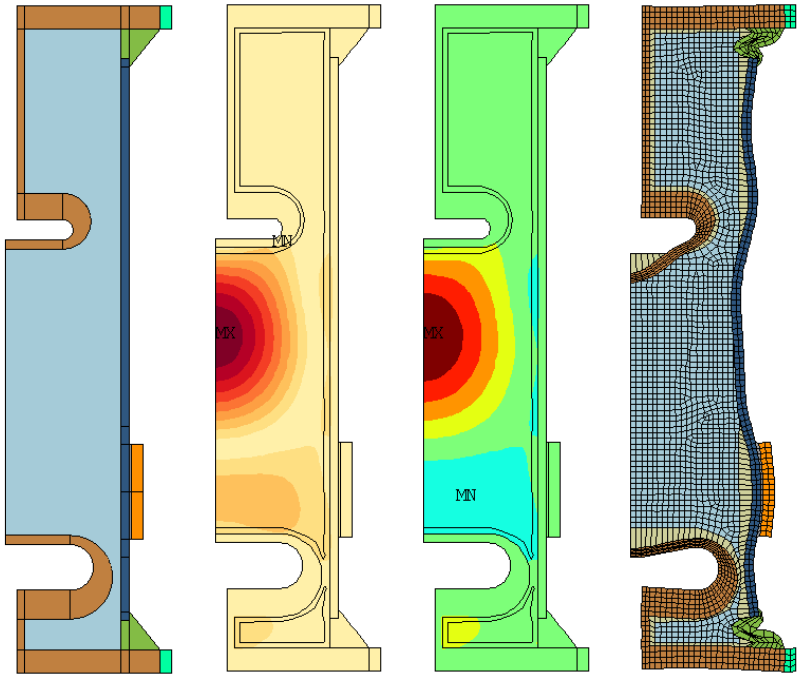


Figure 5.21 Geometry A: case 212
Case 212 represents exactly the same FE model setup and EAO run conditions as case 140 (fig 5.20) with the sole exception that the postprocessing routine calculating the peak interface pressure $p_{\mathrm{if}} = \max(p_{\mathrm{wall}}, p_{\mathrm{p}})$, the wall pressure ratio r_{wp}, and the penalty $\tilde{f}_{\mathrm{pen}}$ was corrected. Not ignoring the piston rims has the effect that r_{wp} is lowered substantially to 22 %. The corresponding changes to the pressure field are visible particularly well in front of the lower piston where the 33 % isocontour (bounding the blue spot in the pressure snapshot of case 140) completely disappears. The change in the upper half is notable upon a close look: here the 33 % isocontour is bounding the orange region which is now not in contact with the piston rim any more. The feature of concave isocontours facing the upper piston is greatly reduced and only remains visible in the 22 % isocontour of the pressure amplitude plot. All this comes however at the cost of a substantially reduced centreline pressure amplitude which dropped from 29 bar in case 140 to 21 bar here.

An interesting question is whether the proposed geometry A can adopt the feature of a mechanical displacement amplification within the glass cylinder. The deformed shape plot shows a wide region of out-of-phase radial displacement stretching from shortly above the transducer all the way up to the height of the upper piston. As the plot does not allow a good quantification or comparison of amplitudes, measured values are given in this case in the resonator specs table. The largest radial displacement amplitude of the glass wall within the height section covered by the transducer is denoted as $u_{r,\mathrm{PZT}}$ and has a value of 3.2 µm. The out-of-phase antinode above it has a much higher amplitude of $u_{r,\mathrm{wall}} = 5.8\,\mu\mathrm{m}$, implying an amplification factor of $A_{\mathrm{wall}} = 1.8$. It is visible in the plot that the normal displacement amplitudes of the piston front plates are still much larger than that, particularly for the upper piston.

(list of parameters: *ih*, *pipos*, *gap_uh*, *gap_lh*, *pt_uh*, *pt_lh*, *pwt1_uh*, *pwt1_lh*, *pwt2_uh*, *pwt2_lh*, *ext_uh*, *ext_lh*, *sh1*, *silext*)

resonator specs	
f_{res}	20 626 Hz
p_{max}	21.12 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	0.81, 2.34, 4.59 bar
p_{wall}	4.57 bar
r_{wp}	0.22
Q_{mech}	390
$u_{r,\mathrm{PZT}}$	3.2 µm
$u_{r,\mathrm{wall}}$ (A_{wall})	5.8 µm (1.8)

optimisation key facts	
EA	THEA
scanned	18-24 kHz
f_{obj}	$max(-Im(p) \cdot \text{sine window})$ with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{p}}, p_{\mathrm{wall}})$
$N_{\mathrm{population}}$	80
generations	80
N_{eval}	6400
local search	downhill-simplex
iterations	100

5.4 Geometry B: simple H-form

Geometry B (fig. 5.22) is in principle just a simplified version of the flange-equipped version of Cancelos' resonator design[6] with H-shaped cross section without the problematic bolted flange connections. Only silicone and epoxy are used for connecting assembly parts. Without much thinking about real-world applicability of this design, the interest of conducting a few EAO runs was just to see what basic vibration mode shapes would emerge. Three optimisation results are shown below representing trials with different materials for manufacturing the piston-endplate parts: aluminium, steel, and glass. Two basic patterns are visible, a short and a long version of the resonator geometry. The two forms were discovered with all three piston materials. The short versions generally yield a maximum sound pressure of around 25 bar, the piston front plate is always quite thin and exhibits a large displacement, or to be more exact, only the central parts of the front plates have large oscillation amplitudes. The long versions achieve values around 21 bar, the front plates are thick, neither front nor end plates move a lot.

Not many optimisation runs were conducted with geometry B. The fact that the piston front plates have only little freedom of motion was identified as the principal shortcoming. Geometry C, discussed in the next section, is a design idea with the aim of alleviating that problem.

[6] i. e. the RPI resonator labelled Nº 8, according to the listing of appendix I.4

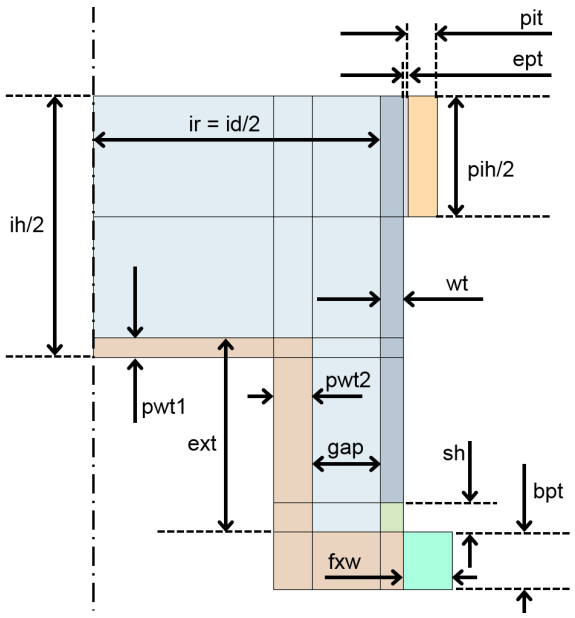


Figure 5.22 Parametrisation of geometry B
In this design the resonator consists of a cylindrical glass tube closed with a top and a bottom head. Only the symmetric case is discussed here where only one half of the resonator needs to be modelled. The head part appears in the 2D axis-symmetric cross section as the simple combination of three rectangles, they are referred to as piston front plate, piston side wall, and endplate or base plate. The heads are glued to the glass tube with silicone, this replaces the bolt connections used in several RPI resonators.

label	description	value
id	inner diameter of main volume	59.2
ih	inner height of main volume	[4,200]
pid	piezo ring thickness	3
pih	piezo ring height	25
ept	epoxy layer thickness	0.5
ext	extension from reflector to cap base	[10,60]
wt	wall thickness of main cylindrical glass wall (hull)	2.4
sh	height of silicone bead sealing the top head	[1,7]
gap	gap between piston rim and inner surface of glass hull	[3,20]
pwt1	piston front wall thickness	[1,16]
pwt2	piston side wall thickness	[1,4]
bpt	base plate thickness	[2,30]
fxw	fixation material width	5

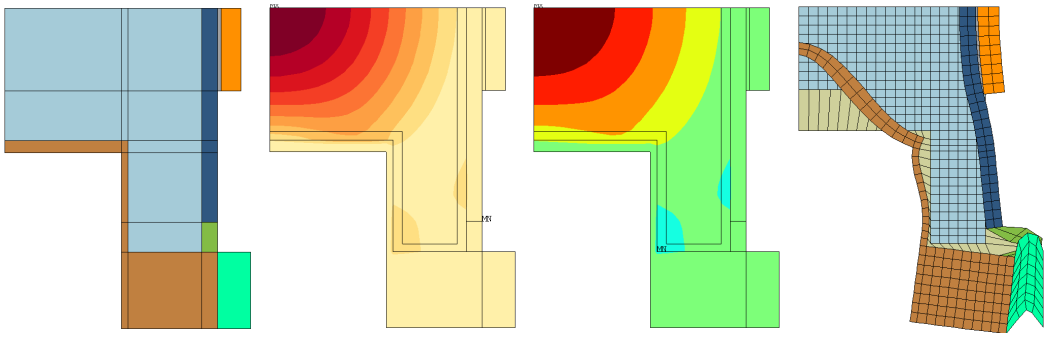


Figure 5.23 Geometry B: case 33

In this case with aluminium as material used for the resonator heads, the optimisation has led to a small liquid volume and a thin-walled piston structure. Just one single antinode fits into the resonator both vertically and radially. The piston front plate has to move a lot in the normal direction in order to serve as low pressure boundary condition. It can fulfil this task only because of a very large displacement amplitude of its central region whereas the outer rim has almost no vertical motion component. The reason is simple. The base plate does not move a lot as it is heavy, and the straight piston side wall, no matter how thin it gets, is still a relatively stiff connection between front and base plate, at least concerning stretching and compression along the vertical axis. This restriction of the freedom for vertical motion of the front plate by the stiff side wall is the problematic aspect of the design. A thin front plate with a lot of deformation is the only way to still enable a strong pressure antinode in the small liquid volume. Moreover, it can be assumed that this setup carries a high potential for fatigue problems, particularly for the edge connecting the front plate with the piston side wall.

(list of parameters: *gap*, *ih*, *ext*, *bpt*, *sh*, *pwt2*, *pwt1*)

resonator specs	
f_{res}	18 944 Hz
p_{max}	26.20 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	3.34, 3.34, 8.24 bar
p_{wall}	3.41 bar
r_{wp}	0.31
Q_{mech}	284

optimisation key facts	
EA	THEA
scanned	15-23 kHz
f_{obj}	p_{max} with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{wall}})$
$N_{\mathrm{population}}$	80
generations	80
N_{eval}	6400
local search	downhill-simplex
iterations	505

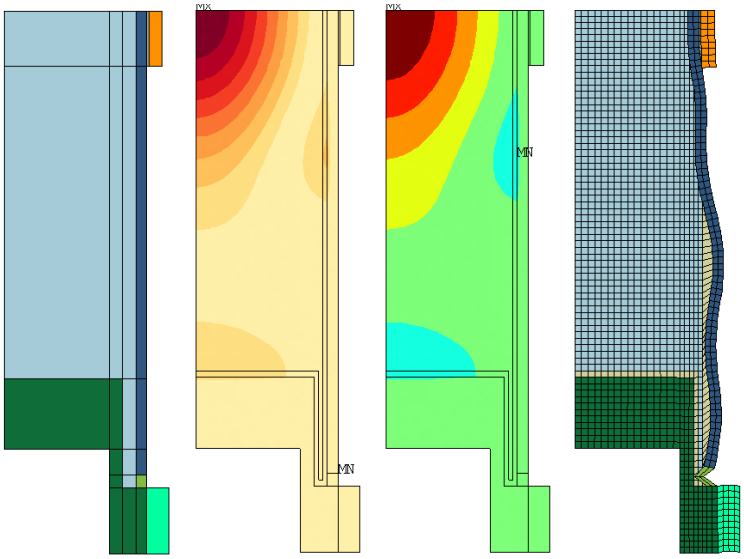


Figure 5.24 Geometry B: case 34

The EAO run with the above result has been conducted under exactly the same conditions as case 33 with the only exception of the piston material which is steel here. With steel pistons this long resonator version was the result of most of the few EAO trials. The piston is a massive part exhibiting even at the resonance frequency a very little motion amplitude. There is neither a lot of bending nor translation. The zero displacement boundary condition imposed by the piston front plates indeed attracts sound pressure antinodes. This is possible without causing a high wall pressure ratio only because the long vertical distance between the two opposing pistons allows the standing acoustic wave pattern to decay far away from the central region in conjunction with a suitable vibration mode of the glass hull.

(list of parameters: *gap*, *ih*, *ext*, *bpt*, *sh*, *pwt2*, *pwt1*)

resonator specs	
f_{res}	18 524 Hz
$p_{\max}$	21.28 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	3.84, 3.84, 3.84 bar
p_{wall}	5.07 bar
r_{wp}	0.24
Q_{mech}	282

optimisation key facts	
EA	THEA
scanned	15-23 kHz
f_{obj}	$p_{\max}$ with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{wall}})$
$N_{\mathrm{population}}$	80
generations	80
N_{eval}	6400
local search	downhill-simplex
iterations	460

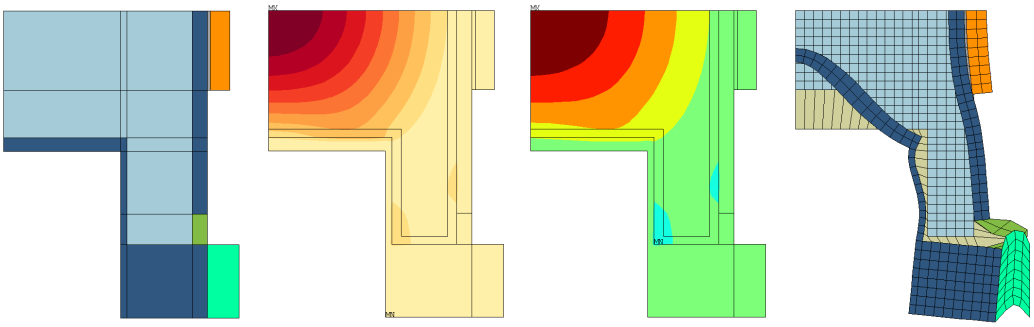


Figure 5.25 Geometry B: case 35
The short optimised version with glass as the piston material looks similar as the aluminium version. The important characteristics $p_{\max}$, f_{res}, and r_{wp} match closely. For the above setup it can be noted that the fine-tuning through local search led to a perfect balance between p_{lpc} and p_{wall} while the largest interface pressure occurs on the outer region of the piston front plate where it was (due to the flawed version of the postprocessing routine) not accounted for during EAO.

(list of parameters: *gap*, *ih*, *ext*, *bpt*, *sh*, *pwt2*, *pwt1*)

resonator specs	
f_{res}	18 968 Hz
$p_{\max}$	26.33 bar
$p_{\text{upc}}, p_{\text{lpc}}, p_{\text{p}}$	3.32, 3.32, 7.79 bar
p_{wall}	3.32 bar
r_{wp}	0.30
Q_{mech}	282

optimisation key facts	
EA	THEA
scanned	15-23 kHz
f_{obj}	$p_{\max}$ with $\tilde{f}_{\text{penC}}(p_{\text{upc}}, p_{\text{lpc}}, p_{\text{wall}})$
$N_{\text{population}}$	80
generations	80
N_{eval}	6400
local search	downhill-simplex
iterations	500

5.5 Geometry C: pistons on flexible discs

The last geometry chosen for EA optimisation was developed with the intention to overcome the problems associated with geometries A & B. The problems of geometry A are the epoxy glue layer behind the piston front plates of glass or the manufacturing challenges of the complicated metal parts in the variants avoiding the glue connection. The problem of geometry B is the stiffness of the piston side wall against vertical stretching and compression.

Geometry C, depicted schematically in figure 5.26, can be seen as a modification of geometry A: the transition can be made by increasing the radius of the piston holding tubes. Yet it can also be understood as a slight complication of geometry B: letting the piston side wall rest on a thin horizontal metal membrane instead of fixing it to the massive base plate directly gives it the needed freedom of vertical motion. At a given thickness of the membrane or ring disc the softness of that bearing can be increased by enlarging the radial distance it bridges. Therefore the model geometry and the parametrisation are adjusted such that the transition between the thin disc and the massive base plate or ring can be pushed outwards as far as possible, even beyond the outer radius of the cylindrical glass hull, by tuning the corresponding parameter. The question is at what point it will become impossible to drill outlets for refilling through the thick part of the metal. In figure 5.27 it is sketched out how the limit can be pushed further away by drilling narrow nozzles at an angle or sideways.

A second question of interest when manufacturing resonators of this design is how to ensure precision and tightness during assembly when making the silicone glue connections. This question is also addressed in figure 5.27. On the right it exhibits a suggested sequence of assembly steps. Laying a silicone bead on a clean surface with a syringe and setting a solid part down on it may not always produce tightness. It is better to first spread thin layers of silicone on the target surfaces (sketch a) because the shear and pressure forces of the spreading help filling grooves on rough surfaces, closing gaps, and enlarging the contact area on smooth surfaces. As a next step, a pre-cured ring of silicone with L-shaped cross section can be pressed down into its place on the end cap (b). When pre-manufacturing such a silicone ring, its dimensions can be better controlled as compared to forming the whole sealing ring from fresh paste. The pre-cured silicone ring can serve as a jig when setting down the glass cylinder (c); it helps to control the distance between the cylinder wall and the end cap. Lastly, the remaining gap on the outside can be filled up with silicone paste or a pre-cured ring can be pushed down into the groove (d).

On the following pages the optimisation results will be presented. Among the possible options (a) symmetric versus asymmetric FE model, (b) single transducer versus two transducers, and (c) aluminium versus steel a restricting choice had to be made. It was decided to investigate only setups with off-centre positions of the transducer where there is a chance of finding working modes exhibiting the feature of displacement amplification in the glass cylinder. This means in the symmetric case the resonator is driven by two transducers. In the asymmetric case only single-transducer setups with the PZT ring in the lower half were examined in order to explore the capability of reproducing the West-Howlett resonator working mode.

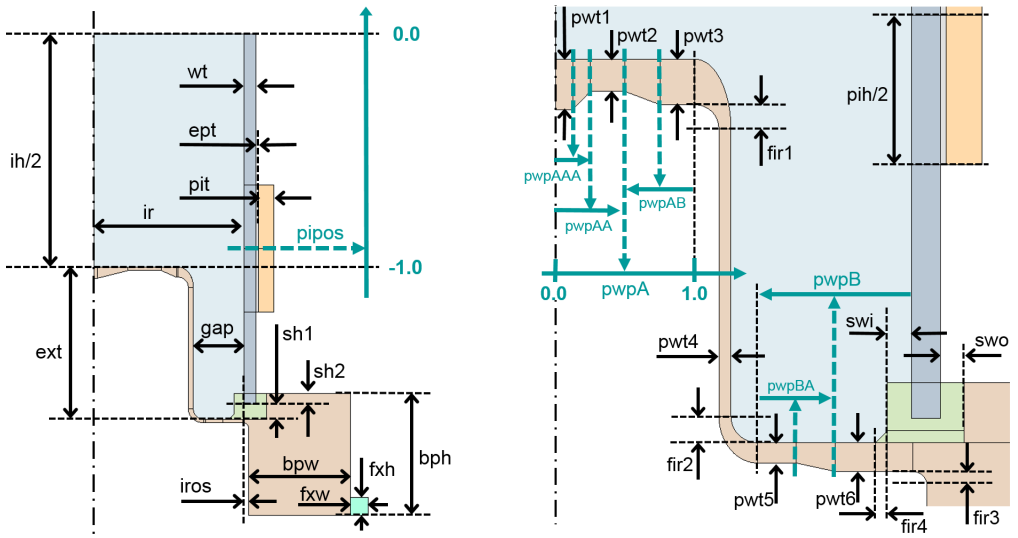


Figure 5.26 Parametrisation of geometry C

In geometry C the cross section of the resonator head part basically has just one more edge than in geometry B. The cylindrical piston side walls do not rest directly on the heavy base plate (or base ring), instead there is one more horizontal plate bridging some radial distance between piston side wall and base ring. This ring disc can be tuned to become quite thin which allows it to act like a drum skin. This is supposed to accomplish a softer bearing and more freedom for vertical motion for the piston front plate and the side wall. In order to allow a combined vibration pattern including internal bending modes of the front plate, sufficient design degrees of freedom are needed. This is the reason why there are six different wall thickness parameters and several accompanying parameters for determining the transition regions. The piston front plate contains three radial segments where different wall thicknesses are possible and the ring disc at the base has two. The remaining thickness parameter applies to the piston side wall. Trapezoidal transition regions connect the radial segments. Owing to the considerations outlined in figure 4.9, the horizontal widths of these segments are not determined directly through distance parameters. Instead, available spaces are divided up by dimensionless parameters. Lastly, it has to be noted that there are ellipsoidal curve segments in this geometry. There are bends at both ends of the piston side wall. Whereas the inner fillet curvatures remain strictly circular, the outer curvatures stem from circle segments being scaled in one direction to account for transitions in wall thickness.

label	description	value
bph	height of base plate (base ring)	[8,50]
bpw	width of base plate (base ring)	[8,50]
epd	epoxy layer thickness	0.5
ext	vertical chamber extension	[12,70]
fir(i)	fillet inner radius (i^{th} fillet)	[0.5,8]
gap	gap between piston rim and inner surface of glass hull	[5,15]
ih	inner height of main liquid volume	[40,120]
ir	inner radius of main cylindrical glass wall (hull)	29.6
iros	inner radius offset of base ring with respect to glass hull	[-10,5]
pit	piezo ring thickness	3
pih	piezo ring height	25
pipos	parameter determining the transducer's vertical position	[-1,-0.5]
pwt(i)	piston wall thickness ($i = 1, ..., 6$)	>0.5
pwpA,pwpAB,...	dimensionless parameters determining division ratios of available space	[0.1;0.9]
sh1	silicone height 1	[0.5;5]
sh2	silicone height 2	[0.5;5]
swi	silicone width inner bead	[0.5;5]
swo	silicone width outer bead	[0.5;5]
wt	wall thickness of main cylindrical glass wall (hull)	2.4
fxw	fixation material width	3.6
fxh	fixation material height	3.6

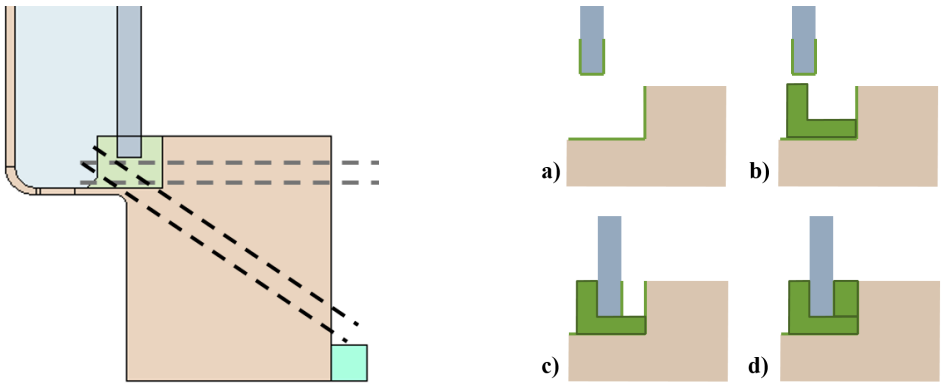


Figure 5.27 Geometry C: feasibility and manufacturing
In the diagram on the left dashed lines indicate where tunnels can be drilled through the metal of the base ring for the construction of outlets in such a way as to avoid too much interference with the thin part which needs to be able to vibrate. The sketches on the right illustrate a basic scheme of sealing the chamber in a way allowing a controlled shape of the silicone beads and control of the positioning of the glass cylinder relative to the metal end caps. The scheme is feasible under the condition that when sealing the second end cap there is no more hand or tool access available to the inner volume of the resonator.

For the symmetric case four optimised results are shown, two with aluminium, two with steel for the metal parts. The difference in EAO setup between the first and second pair is the threshold parameter θ of the penalty function. With a value of $\theta = \frac{1}{2}$ only setups with elevated wall pressure ratios around 30 to 40 % were achieved (see figures 5.28 and 5.29), however with notably high pressure amplitude performance. This first pair of result designs represents short experimental optimisation runs with a high degree of manual interference and non-constant fitness evaluation settings.

During subsequent trials with a much sharpened threshold of $\theta = \frac{1}{5}$ it was possible to decrease the wall pressure ratio of the outcomes to less than 20 % at the cost of a significantly lowered central pressure amplitude. Those two cases are shown in figures 5.32 and 5.33.

In the asymmetric case, where the FE models are twice as large, only very few EAO runs could be conducted. Solely the evaluation routine with mode shape discrimination was employed to search exclusively for setups in a working mode corresponding to the West-Howlett resonator. With the final and flawless fitness evaluation routine two EAO runs were made with aluminium as the piston material and a single run with steel. Two results are shown in figures 5.30 and 5.31.

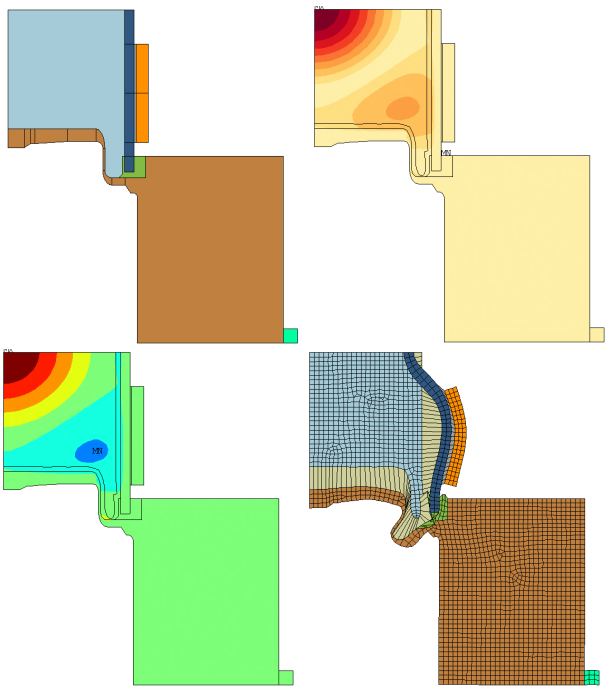


Figure 5.28 Geometry C: case 12

This represents one of the most interesting optimisation results for a symmetric setup with aluminium as piston material. With a sound pressure of 81 bar the resonator exhibits a substantially better performance than any of the optimised setups of geometry A. It is interesting to see that the optimisation has led to a piston front plate which looks very evenly tapered. The thickness of the piston side wall has almost hit the lower boundary of the range from 0.5 to 6 mm allowed for this parameter. An important feature of the found solution is the uniformity of the vertical displacement amplitude across the front plate surface. The plate moves up and down with very little internal bending. The above geometry setup allowed the optimiser to get strikingly close to the performance seen before in figure 5.16 where it had to exploit a meshing error and an unphysical model in order to achieve it. The evenness in structural shape is mirrored on the inside by an evenness in the pressure field. The pressure snapshot shows in yellow-to-red colours the almost spherical shape of the central antinode. It is neighboured by a triangular cyan area outlining the region of tension. Both the pressure snapshot and the pressure amplitude plot show that the isocontours of this triangular region interfere in a perfectly balanced manner with the glass cylinder wall on one side and the piston surface on the other. This is the visible result of the corrected implementation of the penalty computation taking into account the whole piston surface and not only the centre of the piston front. That the local downhill-simplex search has tuned the pressure amplitude peaks on both of these walls towards a close match around 24 bar can be seen in the specs table as well. With a value of 30 % the wall pressure ratio is elevated but not extreme.

This resonator represents the design with the highest pressure amplitude achieved in a valid setup. The performance is based on the one hand on powering a small volume with two transducers and on the other hand on the flexibility given by the thin piston side wall with a thickness of 0.5 mm. It means that besides the elevated wall pressure ratio another shortcoming may exist in a potential fatigue problem of the thin vertical piston side wall connecting the more rigid horizontal pieces.

(free parameters: all parameters indicated as not fixed in figure 5.26 = 25 dimensions)

resonator specs	
f_{res}	25 918 Hz
p_{max}	80.97 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	8.29, .8.29, 24.09 bar
p_{wall}	24.11 bar
r_{wp}	0.30
Q_{mech}	782

optimisation key facts	
EA	THEA
scanned	16-24 kHz
f_{obj}	p_{max} with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{if}})$
θ	0.5 (EAO), 0.4 (LS)
$N_{\mathrm{population}}$	80
generations	30
N_{eval}	2400
local search	downhill-simplex
iterations	960

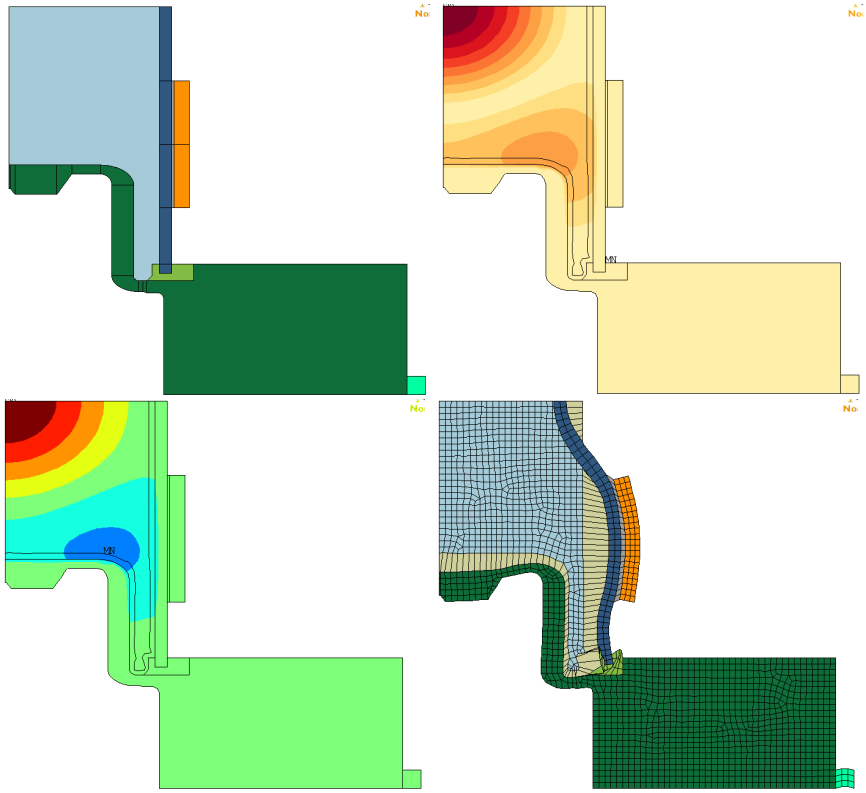


Figure 5.29 Geometry C: case 13
This is the result of the same optimisation setting as in case 12, except that steel was used as the piston and base ring material. After only a short EA run of thirty generations the design represents a less thoroughly optimised setup. The achieved pressure amplitude of 64 bar is in a similar range as seen on other designs with two transducers. In this setup a very inhomogeneous thickness distribution across the piston front with a heavy central weight developed in order to allow for an elevated displacement amplitude in spite of the generally thick steel segments. The low displacement amplitude around the rim of the piston front attracts the pressure antinode seen as blue spot in the snapshot. As a consequence the wall pressure ratio is close to 40 %. This and similar other optimisation results triggered a transition to substantially lowered settings of the penalty function threshold in later EAO trials.

(free parameters: all parameters indicated as not fixed in figure 5.26 = 25 dimensions)

resonator specs	
f_{res}	23 834 Hz
p_{max}	63.96 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	15.29, 15.29, 24.89 bar
p_{wall}	18.20 bar
r_{wp}	0.39
Q_{mech}	592

optimisation key facts	
EA	THEA
scanned	16-24 kHz
f_{obj}	p_{max} with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{if}})$
$N_{\mathrm{population}}$	80
generations	30
N_{eval}	2400
local search	downhill-simplex
iterations	710

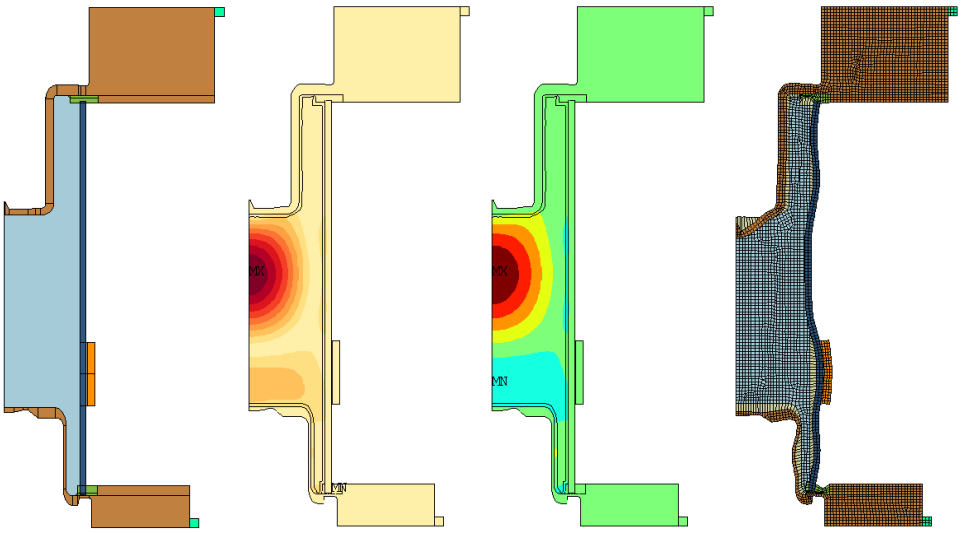


Figure 5.30 Geometry C: case 85

This geometry setup represents one of two full-length EA+LS optimisation runs conducted with the asymmetric model of geometry C. In both cases the mode shape-discriminating evaluation routine yielded the intended pressure field topology with similar performance measures. The peak pressure may seem low with only a little above 20 bar. However, making a comparison with other single-transducer setups of geometry A shows that amplitudes close to 30 bar or above are only achieved in connection with elevated wall pressure ratios. Similarly as seen with geometry A in figure 5.21, reducing the penalty threshold θ to 0.2 has the intended effect of lowering the wall pressure ratio (to around 20 %), but it is accompanied by the side-effect of a lowered sound pressure performance.

The deformed shape plot reveals that the two pistons function quite differently. While on the upper side a front plate with an added central weight bends into a pointed shape, the lower piston moves much more evenly. The flexibility of the lower piston is achieved in two places where the metal thickness becomes thin: along the piston side wall and in the connection to the massive base ring. The motion patterns of the piston front plates have their counterparts in the fluid: near the lower piston the isocontour in the pressure amplitude plot is almost flat whereas in the upper half the amplitude builds up in front of the low-mobility front plate rim which leads to the type of concave contours seen in the West-Howlett resonator and many cases of geometry A. As an effect of the wall pressure-based penalty, however, the amplitude is kept low even at this disadvantageous hot spot.

A question posed at the outset of this work was whether it is possible to achieve a similar displacement amplification effect as in the West-Howlett resonator within the main glass wall of the new designs. The deformed shape plot does not easily reveal it, but the effect exists and can be quantified based on the FEM simulation output data. Corresponding numbers are added to the specs table. At the resonance, the maximal radial displacement amplitude of the glass wall where it is attached to the transducer is $u_{r,\mathrm{PZT}} = 2.8\,\mu\mathrm{m}$. The overall peak of the glass cylinder's radial displacement can be found in the antinode above the transducer where the motion is out of phase. With an amplitude of 5.0 µm the amplification factor A_{wall} is 1.8. Similarly, the piston front plate vertical displacement peaks can be put into relation to the transducer-imposed displacement yielding factors of $A_{\mathrm{lp}} = 1.5$ for the lower and $A_{\mathrm{up}} = 5.5$ for the upper piston.

(free parameters: all parameters indicated as not fixed in figure 5.26 and fully asymmetric yielding 48 dimensions)

resonator specs	
f_{res}	20 966 Hz
p_{max}	20.50 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	0.56, 3.79, 3.86 bar
p_{wall}	3.81 bar
r_{wp}	0.19
Q_{mech}	393
$u_{r,\mathrm{PZT}}$	2.8 µm
$u_{r,\mathrm{wall}}$ (A_{wall})	5.0 µm (1.8)
$u_{z,\mathrm{lp}}$ (A_{lp})	4.2 µm (1.5)
$u_{z,\mathrm{up}}$ (A_{up})	15.4 µm (5.5)

optimisation key facts	
EA	THEA
scanned	18-24 kHz
f_{obj}	p_{max} with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{if}})$
θ	0.2
$N_{\mathrm{population}}$	80
generations	$4 \cdot 14 + 80 = 136$
N_{eval}	10 880
local search	downhill-simplex
iterations	260

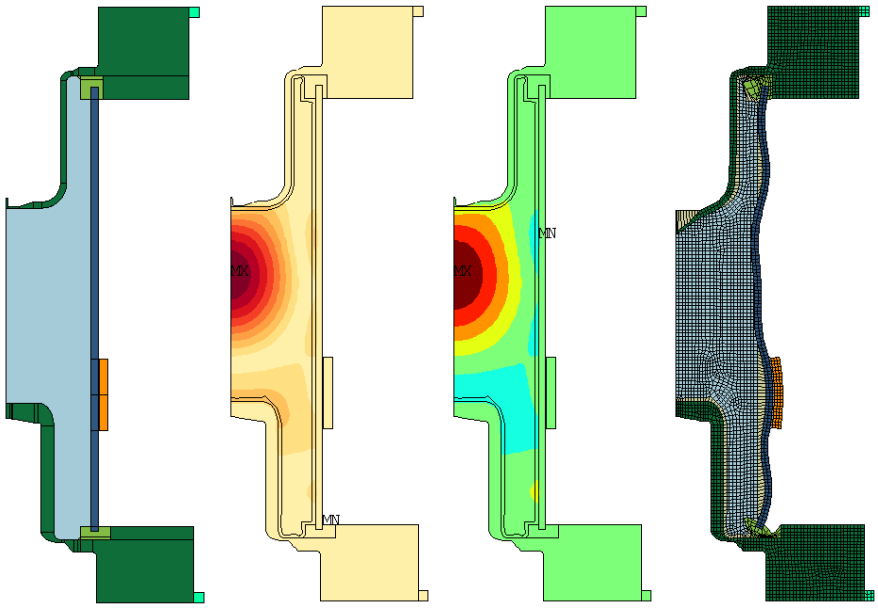


Figure 5.31 Geometry C: case 87

For the asymmetric case with steel only one single full-scale EA+LS optimisation run was conducted. With 19 bar the pressure performance is very similar to its aluminium twin discussed in figure 5.30, only the wall pressure ratio is a little bit worse with 25 %. The upper piston has a thin central region with a little pointed weight in the middle where the displacement is high. The surprising and prominent feature of this tuning result is the sturdy lower piston which exhibits practically no displacement. This is a real exception among all the resonator EAO results. The exceptional shape has its counterpart in the field topology: the isocontours of the weak lower pressure antinode are not encircling a peak at some distance above the piston surface, here they are rather encircling a pressure peak right on the surface. But as it turns out, the pressure amplitude on the glass wall is even higher. The glass wall itself behaves normally and exhibits a twofold higher displacement amplitude on the height of the envisioned cavitation site than below where it is attached to the transducer.

(free parameters: all parameters indicated as not fixed in figure 5.26 and fully asymmetric yielding 48 dimensions)

resonator specs	
f_{res}	20 678 Hz
$p_{\max}$	19.15 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	1.88, 4.69, 4.70 bar
p_{wall}	5.70 bar
r_{wp}	0.25
Q_{mech}	359
$u_{r,\mathrm{PZT}}$	3.0 µm
$u_{r,\mathrm{wall}}$ (A_{wall})	5.6 µm (1.9)
$u_{z,\mathrm{lp}}$ (A_{lp})	0.2 µm (0.07)
$u_{z,\mathrm{up}}$ (A_{up})	15.0 µm (5.1)

optimisation key facts	
EA	THEA
scanned	18-24 kHz
f_{obj}	$p_{\max}$ with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{if}})$
θ	0.2
$N_{\mathrm{population}}$	80
generations	$4 \cdot 14 + 80 = 136$
N_{eval}	10 880
local search	downhill-simplex
iterations	260

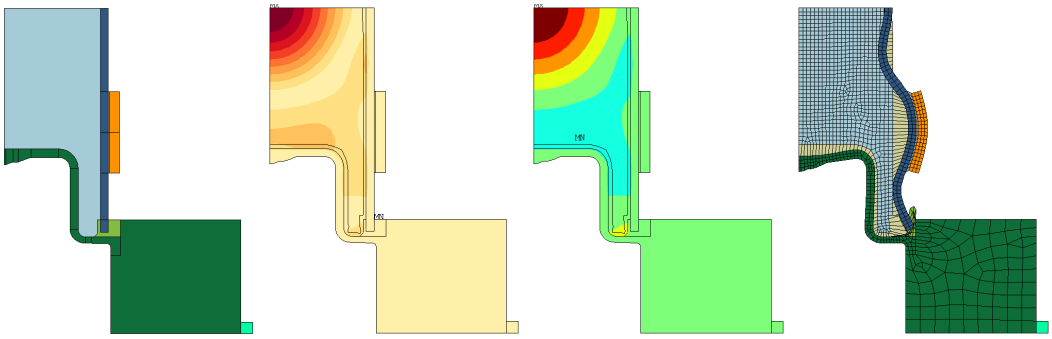


Figure 5.32 Geometry C: case 92

This and the following figure show how the final setup of the EA-based optimisation workflow yielded highly optimised resonator shapes exhibiting pressure field topologies perfectly tailored to fulfil their purpose. In order to increase the reliability of the EA performance, the mutation cooling rate (annealing) was lowered down in these two runs from its standard value of $\gamma = 0.04$ to 0.03. Generally, it can be said that in this type of symmetric setup the weaker in-phase pressure antinode near the transducer serves the shaping of the large, generally more convex, and often almost spherical central out-of-phase pressure peak. Pressure antinodes are regions of converging and diverging fluid motion patterns. Because of this physical origin the isocontours tend to be convex if the interference with irregular surrounding structures is small. Here, the weak antinode has a very broken-up shape. The 11 % isocontour bounding the cyan tension region in the pressure snapshot has lost all tendency towards convexity. When going from the yellow outer rim of the compression region into the tension region, the sphericity of the absolutely dominating central pressure field is extended right into the tension area. Triangular tension regions are apparent in some of the discussed resonators (e. g. figures 5.7, 5.11, 5.17 of geometry A or 5.28 & 5.29 of geometry C), but in no other resonator setup the effect is as strong as seen here.

The motion pattern of the glass wall necessarily plays a role in shaping the internal field topology. This resonator has a larger vertical size as its twin in figure 5.29 allowing for a more complex mode shape in between the transducers. Where in figure 5.29 there is a single out-of-phase displacement peak there is a double peak with a local displacement minimum in the symmetry plane here. The displacement of the double peak is by a factor of 1.4 larger than the peak displacement behind the transducers.

The shape of the piston represents another case where the thorough optimisation has led to a degree of evenness pleasing the nature-trimmed human eye. The cross section exhibits a steady tapering resulting in a moderate added weight in the centre of the front plate. The piston front's displacement is weaker than seen on the transducer, but still in a similar magnitude range. This piston shape with a moderate displacement, a moderate wall thickness, and an even distribution of segment thicknesses appears much less prone to potential material fatigue problems as many of the other optimisation results.

Comparing this design to the analogue EAO run based on a higher penalty threshold in figure 5.29, it can be inferred that the sharper penalty decreases the centreline pressure from 64 to 40 bar and r_{wp} from 40 % to 24 % while elongating the overall resonator shape.

(free parameters: all parameters indicated as not fixed in figure 5.26 = 25 dimensions)

resonator specs	
f_{res}	21 234 Hz
p_{max}	39.57 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	8.14, 8.14, 9.50 bar
p_{wall}	9.49 bar
r_{wp}	0.24
Q_{mech}	356
$u_{r,\mathrm{PZT}}$	6.6 µm
$u_{r,\mathrm{wall}}$ (A_{wall})	8.7 µm (1.4)
$u_{z,\mathrm{lp}}$ (A_{lp})	4.0 µm (0.6)

optimisation key facts	
EA	THEA
scanned	18-24 kHz
f_{obj}	p_{max} with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{if}})$
θ	0.2
$N_{\mathrm{population}}$	80
generations	$4 \cdot 14 + 80 = 136$
N_{eval}	10 880
local search	downhill-simplex
iterations	450

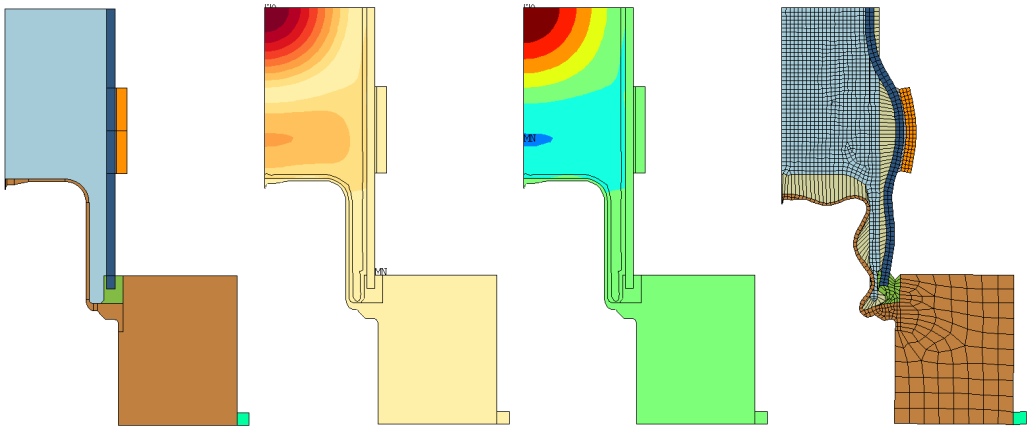


Figure 5.33 Geometry C: case 96
The result of the second one of the final pair of EAO runs is shown, the version with aluminium. Here as well the two 11 % isocontours are bent in the same direction, but the degree of parallelism is not as strong as in the steel version. In this case the harshened penalty forcing a lower wall pressure ratio had very similar effects: the inner volume is elongated in comparison to the less penalised twin in figure 5.28, the pressure performance is reduced to around 40 bar, and a setup with much lowered wall pressure is achieved. In fact, with 14 % it is the lowest wall pressure ratio of all optimisation results.

A particular feature is the thin-walled piston with a small added weight around the central axis. A lot of the piston's flexibility comes from the bellows-like squeezing of the vertical side wall. By consequence, the front plate moves evenly with a large displacement amplitude not only in its centre, but also near the rim which allows for a large tension area visible in the pressure snapshot. A small blue spot indicates the antinode of the tension region. Its position in the middle of the region, far away from the piston surface, or the 22 % isocontour in the pressure amplitude plot which does not touch the piston front plate but runs parallel in front of it are visible indicators of the optimised properties of this sound field topology featuring the lowest achieved wall pressure ratio.

Certainly, the thin-walled but strongly flexing piston raises the question whether there will be fatigue. Future analyses will have to answer this question. Was the lower bound for the wall thickness chosen too low with 0.5 mm? Maybe the one depicted above is not the most practical resonator design, maybe a more practical aluminium version can be achieved constraining the piston wall segment thicknesses with larger lower bounds. Nevertheless, the similar sound fields but strong differences in the piston wall thickness (and Young's modulus) seen in the comparison of figure 5.32 with this one together make a strong message: there is room for further constraints by the engineer, the optimisation routine can tune the structure to achieve the desired sound field either this way or that way.

(free parameters: all parameters indicated as not fixed in figure 5.26 = 25 dimensions)

resonator specs	
f_{res}	21 633 Hz
$p_{\max}$	42.61 bar
$p_{\mathrm{upc}}, p_{\mathrm{lpc}}, p_{\mathrm{p}}$	6.17, 6.17, 6.17 bar
p_{wall}	6.17 bar
r_{wp}	0.14
Q_{mech}	424
$u_{r,\mathrm{PZT}}$	5.8 µm
$u_{r,\mathrm{wall}}$ (A_{wall})	9.0 µm (1.5)
$u_{z,\mathrm{lp}}$ (A_{lp})	12.1 µm (2.1)

optimisation key facts	
EA	THEA
scanned	18-24 kHz
f_{obj}	$p_{\max}$ with $\tilde{f}_{\mathrm{penC}}(p_{\mathrm{if}})$
θ	0.2
$N_{\mathrm{population}}$	80
generations	$1 + 80 = 81$
N_{eval}	6480
local search	downhill-simplex
iterations	720

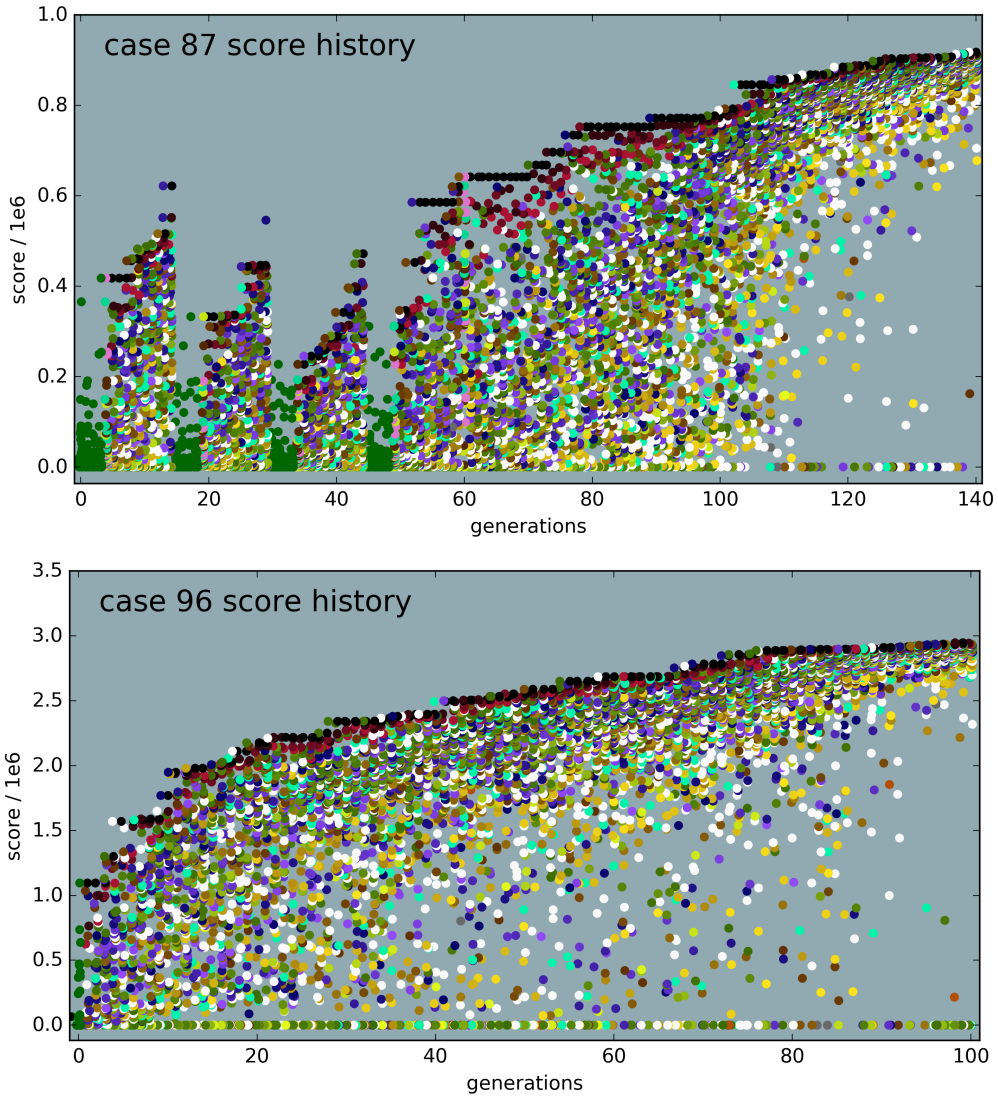


Figure 5.34 Geometry C: Score histories of cases 87 and 96
In the final setup as described and benchmarked in chapter 3 the tier-based hybrid EA produced these exemplary score history diagrams when applied to the resonator optimisation task. Case 87 (top) shows the four consecutive random initialisation phases (green dots) when the population merging scheme is turned on. The lower plot shows that case 96 was a simple run without the population merging scheme. When comparing these plots to the earlier conducted EA run plotted in figure 5.12, two differences are clearly visible: on the one hand there are no negative scores due to the changed penalty computation and on the other hand there are more colours because more tiers were added to the hybrid EA scheme. Cyan dots show that the second crossover operator, cigar-CO, has been implemented (see colour legend in figure 3.6). Yellow-green dots represent chromosomes created in DE manner. Looking at the improvement steps forming the upper edge, it can be seen that there are no extended stagnation phases and that progress is made through chromosomes with diverse tier origin.

5.6 Summarising the resonator EAO case studies

Pre-existing geometries

Four old and new SF resonator geometries were investigated as target of optimisation with evolutionary algorithms. The oldest resonator design is the one by West & Howlett which was used by Taleyarkhan et al. and Saglime et al. for SF trials. The two corresponding optimisation case studies serve as a proof of concept and a proof of efficiency for the EAO approach. They show that the automated blackbox optimisation approach may require a somewhat increased computational budget but allows to avoid and is superior than lengthy investigations in the form of multi-step parameter studies requiring repeated human interference. In a preceding study (documented in appendix Q) the investigation of this geometry brought about the proof of necessity for a global optimisation approach which is robust against being trapped in local optima.

The other pre-existing resonator design is the resonator with H-shaped internal volume developed by Cancelos. It is investigated in a slightly modified and simplified geometry with metal or glass end caps connected with silicone to the cylindrical outer glass wall, which is labelled geometry B herein. A close look at the results of few optimisation runs reveals as a crucial problem the too stiff connection between the two halves of the piston, the front plate and the outer ring. The piston shape is too simple and does not allow the degrees of freedom for an internal vibration mode of the front plate swinging vertically against the ring. The optimisation algorithms achieve decent sound pressure performance only by making the piston front plate membrane-like thin or by resorting to lengthy resonator shapes. The elongated setup seems to permit sound fields with a relatively low wall pressure ratio in spite of heavy pistons exhibiting very little displacement and furnishing a boundary condition close to the trivial hard wall. If the pistons can thus be taken out of the equation then only the geometry of the glass cylinder with the transducer on it remains as the sound field-determining structure. This observation is noteworthy, as sturdy "hard wall" pistons would solve practical problems: they are easy to manufacture, they can serve as chamber mounting points, they offer unproblematic locations for outlets.

Besides the two old geometries (the West-Howlett resonator and geometry B) two new geometries (A & C) were investigated and are suggested as new solutions to the SF resonator design problem overcoming the limitations of the old ones. For both geometries it is shown that the EAO approach can find vibration mode shapes yielding comparable or better performance in terms of $p_{\max}$ (to be considered together with r_{wp}) than the West-Howlett design in a manually optimised setup which serves as the baseline for all EAO case studies.

Geometry A as learning case: technical aspects of SF resonator tuning by global and local search and aspects of general interest in the context of applied blackbox optimisation

Geometry A served as the learning case for the application of evolutionary algorithms as global blackbox optimisers. Many different case studies are presented with the intention of not only highlighting the outcome in terms of SF resonator propos-

als but also insights of general interest in the context of applying EAs and other blackbox optimisers to real-world problems. Given a robust model parametrisation and many degrees of freedom EA search can be used as a powerful tool for design space exploration. This is illustrated by the many different mode shapes (sound field topologies) generated by differently proportioned realisations of one and the same base geometry. Enchaining global EA with local simplex search allows a quick cycle of exploration, discovery, and comparison. Thoroughly optimised exemplars make different resonator design concepts or different working modes of one design idea comparable.

However, there are issues requiring attention ranging from the objective function definition over simulation data postprocessing to search space topology and constraint enforcement approaches. For geometry A an extended range of case studies is presented including low-quality and undesired results of particular optimisation runs. These cases complement and illustrate the theoretical and technical descriptions of the attention-requiring issues compiled in the preceding chapters and render this documentation of the EA-based resonator optimisation approach complete. The documentation would be incomplete without consideration of such technical aspects, and the gap would raise the hurdle for successfully taking up and continuing the presented work by others.

Explorative versus targeted usage of global search algorithms

It has been pointed out that the global EA search approach allows to explore many different possible sound field topologies. If more parameters are opened up for tuning, if more degrees of freedom are made available, then more diverse solutions can be discovered. By contrast, it is also shown that in a modified setup with as many degrees of freedom on the side of the input variables but heavy filtering and discrimination on the side of the postprocessing the EA+LS search allows a fast and robust method of tuning different FE model variants towards a predetermined and intended working mode. The last case studies of geometry A document the efficiency of the targeted usage (as opposed to the explorative usage) of the global search algorithms. One case study based on pure local search proves the (still existing) necessity of the global search approach for the application. In single-objective optimisation the objective function is one scalar number representing the quality of an evaluated solution candidate. The postprocessing steps for distilling such a number can be simple (e. g. collecting the maximum of a physical quantity like pressure) or more complicated. The objective function interfaces the optimisation target and the optimiser while representing the intention behind the whole optimisation procedure. Decisions on data postprocessing, filtering, and constraint penalisation can have just as heavy an effect on the efficiency of the optimisation procedure as decisions concerning the choice and setup of algorithms. The case studies show the effect of employing the evaluation schemes based on wide followed by narrow frequency sweeps described in the preceding chapter. Based on a simple evaluation scheme of searching for the maximum pressure and penalising high wall pressure ratios different resonator working modes (i. e. mode shapes) can be explored, whereas based on mode shape discrimination a targeted optimisation can be pursued instead of exploration.

New SF resonator design proposals: geometry A

Primarily, however, the EAO case studies of geometry A have the goal to answer the question of suitability and practicality of this resonator design for SF experiments. The suitability question can be answered with a clear yes as the design allows for the adoption of the same mode shape and the same beneficial mechanical displacement amplification mechanism as the West-Howlett design while at the same time overcoming the reproducibility problem because of allowing precision machining of the structural components. The practicality question can only be answered with a conditional yes until further experimental trials prove the reliable manufacturing and usage of the two proposed piston versions: the glass-metal composite version requires testing for durability, the all-metal version requires establishing a way of producing the hollow shapes, e. g. through laser welding, friction welding, or 3D printing with subsequent surface treatment.

New SF resonator design proposals: geometry C

Geometry C represents an SF resonator design of which the parts can be manufactured with conventional metal machining techniques. The geometry was developed in the search for a solution to the problem hampering the piston front plate motion in geometry B. Additional edges in the piston cross section create the additional degrees of freedom allowing vertical displacement of the central part of the piston while keeping the massive outer ring (the bearing and counterweight) at rest. The presented case studies show how the EA-based optimisation approach allows to get in a straightforward manner from the design concept over the parametrised FE model and the EA+LS tuning procedure to thoroughly optimised and well-performing resonator setups. Resonator geometry C also allows for hosting the sound pressure mode shapes of the West-Howlett design while enabling a mechanical amplification of the glass hull motion with respect to the transducer. Highly optimised symmetrical setups with two transducers are presented as well.

With geometries A & C as new proposed SF resonator designs and having established a powerful framework of algorithmic resonator tuning this work makes a decisive contribution to research around sonofusion. It shows the way how the next generation of SF experiments can be conducted employing performant and at the same time reproducible resonators. The prospect of systematic reproducibility adds a crucial justifying argument to future attempts of retrying SF experiments in the setup introduced by Taleyarkhan et al., a backing which is currently lacking.

5.6.1 A note on the question of material choice

Two metals with very different properties, aluminium and steel, were considered besides glass as primary structural materials for the acoustic resonators. Aluminium as piston material had been favoured over steel by Cancelos when designing the H-shaped resonator with flanges[7] because of the aim of matching the bulk acoustic impedance $Z = \varrho c$ of the liquid more closely [69]. Looking at the different optimised resonators presented here, one can see steel winning by $p_{\max}$ in the comparison

[7] resonator Nº 8 in the listing of appendix I.4

between figures 5.17 and 5.20 (geometry A) and aluminium taking the lead among all three pair comparisons of geometry C. But the number of pairs of EAO runs is too low, the working mode within any pair too different, and the material-independent damping setting too unrealistic to allow a conclusive judgement. In any case it should be kept in mind that the acoustic impedance at a reference plane is generally defined as the complex ratio between pressure and volume flow $Z = p/\Phi$. So, the acoustic impedance of a structural surface is highly determined by its vibration response. By tuning wall thicknesses in various places to control the masses and stiffnesses in the structure, various kinds of impedance patterns can be achieved with any kind of material. A comfortable aspect of an EA-tuned FE model is that these features are addressed implicitly (maximising sound pressure and minimising or limiting the wall pressure ratio will automatically lead to the achievement of suitable impedance boundary conditions).

5.6.2 Limitations of the resonator tuning study

The greatest weakness of the present study lies in the material data library underlying the FE models. This owes to the fact that the experimentally characterised resonators[8] do not represent optimal cases for model validation and material data calibration.[9] Currently, the FEM simulation results for the reference resonator[10] of the West Howlett type do not perfectly match the benchmarking data on sound pressure, displacement amplitudes, and electrical properties gathered by lab measurements. Thus it can be inferred, that upon building one of the proposed optimised geometries, the predicted acoustic properties will also not be matched exactly. It is very likely that a real-world exemplar will land in a working point somewhat off the anticipated resonance. Ideally, the intention of this work would be to publish a detailed plan with exact dimensions for constructing the next resonator to use in SF experiments. But this is not possible so far, and the statement at the end of this project has to be reduced to the following: *New resonator designs are being proposed. These designs will be suitable for SF experiments after a slight update: their dimensions should be locally optimised again after having made available better material data and having gained the ability to match lab-measured resonator*[11] *characteristics with FE models.*

The technical reasons for this shortcoming should be remembered, they are threefold: (a) literature data on material properties is insufficient, in particular when material properties depend on the production history (e. g. conditions of polymerisation of epoxy and silicone, of casting, annealing, and cold deformation of metals, of crystal structure formation in piezoceramics), (b) model calibration is impossible when too much needs to be calibrated at the same time (this is why unreliable outcomes are to be expected if the calibration of all unknowns is attempted based on characterisation data from a single already assembled West-Howlett resonator because it

[8] denoted as resonator Nº 5 & 8 according to the listing in appendix I.4

[9] See appendices O & Q describing the preceding efforts for experimental resonator characterisation and FE model setup and calibration.

[10] i. e. resonator Nº 5

[11] A collection of resonators for that purpose should comprise besides complex geometries simple study objects like metal plates, glass cylinder segments, free transducer rings and so on.

involves too many structural components, too many materials, and too many geometry details and dimensions not known with sufficient precision), and finally, (c) a 2D-axis-symmetric FE model cannot match a resonator exhibiting deviations from axis-symmetry.

Furthermore, all optimisation runs presented in this chapter involve the simplification of one global damping ratio instead of material-dependent damping constants. Secondly, all these FEM simulations were conducted with a preliminary and less realistic dataset for the properties of the piezoceramic (see appendix Q.1.2, particularly table Q.1, p. 423). This was done to ensure the comparability of the whole set of optimisation results.

It is nevertheless assumed that the transition to a future updated material library will not invalidate the central insights and statements presented herein: the same resonator designs will be able to exhibit the same vibration mode shapes and sound pressure field topologies, only the structure proportions will require local fine-tuning and resonance patterns might morph accordingly. Yet, the principle character of the resonator design problem stays unchanged together with the suitability of the presented approach to address it.

Another limitation of a different kind of the present study is the fact that optimised designs are not compared under the aspect of robustness of the local optima. It is clearly a topic where future studies will be able to add insights. The present study goes the most pressing step forward towards robust resonator designs by guiding the way from luck-determined to systematically laid-out precision-machined resonator designs. Going from design optimisation to robust design optimisation (RDO) [388] is the next logic step.

5.6.3 Addressing a view of scepticism: trying to solve engineering problems with random-based optimisers instead of critical thinking is lazy and inefficient

A criticism of using EAs or other black-box optimisers in the search for better solutions to a real-world problem could be put forward along these terms: black-box optimisation is a mentally lazy strategy trying to avoid the effort necessary to fully understand the behaviour of the target system; some low-hanging fruit may be harvested with this lazy blind search while huge computational resources will be consumed and human effort wasted on coding and debugging; but after the straw fire goes out the lack of problem understanding will backfire. In that context, the presented case study adds to existing literature on EA applications to real-world problems. This public library of case studies can help other researchers and engineers to assess the validity of the scepticism described above under their own circumstances and to balance it as realistically as possible against the advantages which can be expected from the incorporation of black-box optimisers and eventually EAs into the own toolbox of skills. It has been tried to make this project documentation a valuable contribution also in this context. While appendix T serves as a general overview and motivation of EA usage, the above chapter underlines some particular experiences gathered with this EA application case of resonator optimisation. On the one hand the issue of the necessity of robust simulation calls was described as well as

the learning curve and development steps mirrored in different versions of the trial evaluation routine. These points can be made into arguments on the EA-sceptic side. On the other hand, it has been tried to document thoroughly under which conditions the EA approach to resonator tuning became a successful case of EA application. Occasionally, it has been tried to point out that the addition of EAO to an existing problem treatment approach does not conflict with efforts towards principled problem understanding and the goal to make global blackbox optimisation obsolete at a later point. In that respect, the above sceptic standpoint is wrong. One the one hand, it could be exemplarily shown, that the optimisation runs themselves yield various pieces of feedback information enlarging the knowledge base about the characteristics of the optimisation problem. But as a more important point, the diversity of the different local optima found by the global search offer valuable input for triggering new thinking processes. The presented series of different optimised pressure fields can illustrate this point. Surely not all of these modes would have been found by manual human experimentation with the FE models of the new geometries. Particularly, modes not fitting into the schematic of n antinodes in the axial and m along the radial direction may represent less straightforward ideas and improbable discoveries during aimed human search. Upon discovery of new resonator modes, the analysis of the structural vibration allows to identify the essential features, and this in turn allows for a next step of thinking in terms of geometry design and considering alternative structures which could host similar or better pressure fields but would be more advantageous in terms of manufacturing or other technical aspects. The experiences made with EAO in this project support the view that when the real-world problem is hard enough, adding EA optimisation to the toolbox of treatment techniques is able to accelerate the speed of progress and the iteration loop between understanding and invention.

In Summary, the general view can be supported that resorting to black-box optimisation tools is not a desperate approach with the aim to avoid problem understanding, but should be seen as a practice with the potential to broaden the thinking about a design task and to trigger deeper problem understanding while ensuring fast turnaround cycles of reliable and efficient parametric optimisation workflows.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
$A_{\text{wall}}, A_{\text{lp}}, A_{\text{up}}$	displacement amplification of wall, lower, upper piston
a_i	lower bound of search space along i^{th} dimension
b_i	upper bound of search space along i^{th} dimension
c	speed of sound/light
f_{obj}	objective function
f_{res}	resonance frequency
f_{pen}	penalty function, contribution to an objective function
$\tilde{f}_{\text{pen}}$	transformed penalty function projecting from $\mathbb{R}$ onto $[0, 1]$
N_{eval}	number of design evaluations
$N_{\text{population}}$	population size (algorithm control/setup)
P	probability
p	sound pressure amplitude
p_0	static pressure
p_{max}	sound pressure maximum throughout the resonator
p_{p}	peak sound pressure detected along entire piston interface
$p_{\text{lpc}}, p_{\text{upc}}$	peak sound pressure probed at centre of lower/upper piston
p_{wall}	peak sound pressure detected along inside of cylindrical glass wall
Q	quality ("pointedness" of a resonance peak)
Q_{mech}	mechanical Q-factor
$\mathbb{R}$	real numbers
r_{wp}	wall pressure ratio, a measure for judging resonator design quality
u	displacement
u_r, u_z	radial, axial displacement
x_i	design parameters
Z	impedance

List of Greek quantity symbols

Symbol	Description
γ	annealing factor (mutation step size reduction)
ζ	damping ratio
θ	penalty function threshold parameter
ϱ	density
Φ	volume flow

List of abbreviations

Abbreviation	Description
APDL	Ansys Parametric Design Language
BC	boundary condition
CMA-ES	evolution strategy with covariance matrix adaptation
CO	crossing-over, crossover
DE	differential evolution
DOF	degree of freedom
EA	evolutionary algorithm
EAO	evolutionary algorithm optimisation (meaning optimisation by evolutionary algorithm)
FE,FEM	finite element (method)
FSI	fluid-structure interaction
KIT	Karlsruhe Institute of Technology (Karlsruher Institut für Technologie)
lpc	lower piston centre
LS	local search
PZT	lead zirconate titanate (a piezoelectric ceramic)
RPI	Rensselaer Polytechnic Institute
RTV	room temperature-vulcanising (silicone)
SF	sonofusion
THEA	tier-based hybrid evolutionary algorithm
upc	upper piston centre

Chapter 6

Conclusion & outlook

The present study was first and foremost aimed at investigating the controversy about SF experiments on an aspect not treated with the appropriate attention in past discussions, namely the aspect of the properties and the general reproducibility of the acoustic resonators. In order to shed more light on the issue, finite element models had been created, benchmarked, and used for a sensitivity study in a preceding work. This work adds the consequent next step in leveraging the FE modelling capability for the simulation-aided design of improved resonators for future SF experiments.

Besides optimised resonator geometry instances the optimisation algorithm-based design methodology itself is considered to be a key contribution to the fields of SF research and resonator engineering in general. A new class of test problems and a hybrid evolutionary algorithm are considered to be valuable contributions to the field of evolutionary computation.

Below, the key insights of this study are first outlined in short form, then in more detail, and finally rounded up by a few remarks of contextualisation.

6.1 The main insights

6.1.1 Key implications on the SF controversy

To solve the SF resonator design problem, an EA-based approach of FE model optimisation was motivated, its application issues discussed, and the results presented. What can be learnt from this project documentation is that

1. an efficient global search tool is needed for the resonator tuning problem,

2. without simulation and an efficient global optimisation scheme resonator designs cannot be properly evaluated and compared, and that

3. it is possible to systematically improve the acoustic resonators, firstly, by transitioning to manufacturing procedures allowing greatly reduced tolerances and, secondly, by reliably realising an optimal or near-optimal sound pressure amplitude at the centre of the liquid volume through parameter tuning of calibrated FE models with global optimisation algorithms.

4. In the context of experimental SF trials, it means that the reproducibility issue (with respect to the resonators) can be resolved by switching from not optimised low-precision resonator assemblies to optimised high-precision and robust designs. The data collected with the FE models of the proposed new geometries (geometries A & C) and the comparison to FEM simulations of the original West-Howlett (i. e., ORNL) design is highly indicative that this transition can be made without reducing resonator performance.

6.1.2 Proposing an EA-based approach of SF resonator design

A recapitulation in more detail of the design task and the EA-based solution approach will allow to enhance the clarity of the above conclusions. Based on experience with SF resonators gathered at RPI during experimental campaigns and based on experience with 2D FEM simulations conducted at RPI and KIT, the SF resonator design problem could be outlined in a comprehensive way. The emergence of an optimal acoustic field depends not only on the shape of the liquid volume, but also on the vibration behaviour of the bounding structure. SF resonator design means deciding on the topology of the sound field and the shape of the enabling structure to contain and excite the liquid volume. The structure has to be able to vibrate in a suitable manner, damping and heat generation have to stay within limits, and at the same time an eye has to be kept on all the other practical aspects like manufacturability and assembly techniques, cooling, refilling, the choice of transducer and its mounting and so on. Due to the "competition of many resonances" and their amplitude growth and decrease under the variation of design parameters the acoustic pressure performance turns out to be highly sensitive to geometry and proportions. The resonator design problem can be dissected into two parts, the actual design task aiming at finding a geometry able to negotiate the requirements of the vibration mode with the ones due to manufacturability and nozzle placement as the first, and the proportion tuning as the second part. Building on the experiences made with the FE model of the ORNL resonator the parameter tuning task could be described as a hard nonseparable optimisation problem. This characterisation could be verified later while tuning new geometries.

In front of a motivational background given in appendix chapter T for using EAs for solving such optimisation tasks, the systematic choice and setup of appropriate EA schemes was described. The choice of which EA to apply is an important one, and it was made here based on a deliberate selection of test problems for benchmarking a selection of algorithms. CMA-ES has been chosen from the literature, and at the same time an new hybrid EA scheme has been devised which exhibits comparable performance on the test problems. The issue of interfacing the algorithm with the simulation and the central role of the objective function (fitness function) has been described in detail. Through the documentation of various optimisation results it could be shown how different objective functions lead to the development of resonator setups with different features. One could say that the fitness function is not only a small subproblem, it is equally important as the optimisation algorithm setup. With more sophistication (i. a. filtering) on the side of the solution evaluation routine the two EAs could be forced to find only a targeted, predetermined mode

shape. A good imagination helping to understand that point is to think of attractor regions of certain solution types in the search space. When devising the solution candidate evaluation subroutine, one can decide to amplify attractors with desired features while damping others. The presented optimised versions of geometry A further illustrate this: many different vibration modes and pressure field topologies could be found with a non-discriminative objective function. But when mode shape-discriminating penalisation schemes were incorporated in the fitness function, the EAs could be compelled to ignore all other mode shapes and converge consistently only on setups exhibiting the intended working mode. Thus, a distinction between a targeted and a more explorative usage of the global search algorithms could be demonstrated. The targeted usage surely has the potential to reduce the number of wasteful EAO runs for achieving a predefined design goal. However, the powerful tool of fitness function sophistication should be used carefully and only after deliberate decisions, for not everything can be obtained at the same time: there is a goal conflict between exploration and efficiency.[1] Heavy discrimination of mode shapes (or other solution candidate features) can yield a slim, efficient, and robust global optimisation procedure geared at tuning towards a well-defined intended setup; but a simpler and freer optimisation approach can be more beneficial for exploring the design space, for learning about the wide range of working modes possible with a new resonator design concept.

In the presented study of SF resonator FE models the explorative as well as the targeted approach of the EA-based design optimisation were used, the explorative approach in order to see what is possible in the design space of a new geometry idea, and the targeted EA-optimisation setup for quickly optimising new geometry variants in order to make them comparable. The global EA search proved to be a central tool of addressing the SF resonator design problem. The simulation-aided and EA optimisation-based way of exploring and comparing resonator design ideas represents a qualitative step forward and introduces a systematic design approach not yet seen in field of SL and SF resonator development. To sum it up in one sentence: the evaluation of one geometry instance is not the same as the evaluation of a design idea, and EA optimisation makes the latter possible.

6.1.3 The development of an efficient hybrid EA

The efficiency of EAs on challenging optimisation test problems can easily be measured to great accuracy, but *why and how exactly* they work can generally not[2] be described with satisfactory precision. There is no "physics of EAs" which would make a stringent connection between the "microscopic level" of mutation and recombination operators and the "macroscopic level" of the global search properties. Nevertheless, the field of evolutionary algorithms (or broader: evolutionary computation) represents a valuable modern addition to the scientific landscape. The first reason is that EC allows to deepen our understanding of the theory of evolution by empirical research, it is a great enhancement to the otherwise much smaller window

[1] That goal conflict is a manifestation of the *no free lunch (NFL)* theorem (see appendix T.1.4).

[2] with the exception of particular algorithms, e. g. ES, which are conceptually very simple as to allow a thorough mathematical analysis in terms of probability distributions

given by pure biological research. The second reason is the general applicability of EAs to so many parametric optimisation problems in engineering and elsewhere which are also part of the natural world.

In parallel to evaluating state-of-the-art EAs on a selection of test problems for their suitability for application to SF resonator optimisation a new hybrid EA scheme was developed: the tier-based hybrid EA (THEA). It is based on dividing the parent and offspring population up into segments or tiers and combining classic EA ideas on the level of mutation and recombination operators. The offspring population segments do not only differ in terms of chromosome generation routines but also in terms of parent selection where there is a combination of rules drawing from classic EA descriptions. Across the five selected test functions THEA exhibits a very competitive performance and even has a slight advantage in terms of robustness. Low sensitivity towards problem properties within a given class or selection of tasks is a valuable aspect when searching the right black-box optimiser for a real-world task.

THEA furnishes an easy-to-use testing lab for EA developers interested in examining EA blends. Its architecture invites for further extension by adding in more basic EA schemes. The EA blending ratios can be seen as degrees of freedom for tuning this hybrid EA, but more importantly, this flexibility allows one to empirically investigate the dependency of the performance on the mixture. While recent EA literature documents wide-spread efforts of singling out the most efficient ingredient from a combination for the purpose of devising adaptation schemes or within a selection-from-portfolio approach as the extreme consequence, segment-based EA hybridisation puts the focus on the question: if and when does the combination work better than the single ingredients? Evaluating statistics while sweeping through blending ratios allows to measure when this is the case.

6.1.4 The development of a new class of visualisable optimisation test problems

Test problems are essential to developing optimisation algorithms. Some problems are trivial but are appreciated for their ability to check whether a new algorithm is robustly working and not prone to known particular types of weaknesses. Some test problems are made to be more challenging e. g. by reflecting the curse of dimensionality. The hardest test problems are *deceptive functions*, where in the extreme case the best search strategy is not to rely on collected data any more, i. e. pure random search. It is easy to devise test problems of the too easy or the extremely hard type, but the relevant range is somewhere in the middle because optimisation problems encountered in nature[3] exhibit objective functions with some level of interpretable structure, functions which are not deceptive throughout.

[3]This has to do with the compressibility of engineering problems or problems encountered in nature in general (see appendix T.1.4): A structure-less objective function in a high-dimensional space must have an exponentially growing number of local valleys of different widths and shapes. The description of such a landscape requires a large amount of data. Problems of which the geometry and the governing equations can be given within a few kilobyte of information cannot give rise to objective functions of unrestricted complexity (leaving aside deterministic chaotic systems like a magnet pendulum).

While searching for an EA suitable for resonator optimisation, a test problem was developed with similar[4] key properties on the one hand, and being intuitively visualisable one the other. A detailed description of the test problem and the generalisable problem class was published in a conference paper [428] and appears here in excerpts as appendix chapter V.2 because it is not part of the direct chain of thoughts leading from the SF resonator design problem statement to the EA-based solution approach. However, since it is deemed to be a non-negligible contribution to the EC community, it will be mentioned here briefly.

The *charged marble problem* is an optimisation test problem based on the task of minimising the sum of two types of potential energies of "repulsive marbles on a hilly ring track", so the two types of potentials are the inter-particle potential and the particle-on-track potential. With a final interest in SF resonator tuning, the interesting and relevant features of the charged marble problem are the diverging energy barriers between local minima and the incorporation of a combinatorial aspect into a real-domain search problem. Besides that, it has several beneficial properties seldomly manifest together in other test problems: it is hard, it is visualisable, the visualisation is understandable intuitively, the aspects making the problem hard can be switched on one by one (e. g. weight difference for making the particles non-identical or local minima on the ring track), smooth hills can be replaced by other structures, and switching from single-objective to bi-objective treatment is directly possible. Even a low number of more objectives can be added. All these properties make the charged marble problem extremely useful for anyone working on improving an optimiser, but particularly for EA development where creative invention and experimentation are such substantial ingredients. A challenging test problem for on-the-fly evaluation of algorithm modifications cannot replace sound statistical tests on a telling suite of benchmarking problems, but it can allow for more rapid prototyping in algorithm development, thus increasing the room for freedom and creativity.

6.1.5 Sorting conceptual levels

From the proposal of specific SF resonator designs to a test problem class for rapid EA prototyping, the contributions made by this work span several conceptual levels. On the lowest level particular instances of new resonator geometries are described. On the next level the EA optimisation approach furnishes a methodology for tuning new design variants and bringing them into the collection of comparable FE models. Above that is the level of different tuning approaches. Here the described hybrid EA allows one to incorporate additional segments in the future and the discussed optimiser evaluation framework allows one to take these extensions or completely different algorithms into consideration and to create a state of comparability in between the optimisers. Another different conceptual level is the one of general thoughts about EAs and their working principles where it is sought to gain understanding and invent ever better EA ideas. On this level the present work provides a comprehensive overview (apdx. T) of EA ideas, motivations, and observations, and

[4] i. e. representative for the SF resonator tuning task where the local optima arise from the "competition and interaction of many resonances" and the steady variation of the associated mode shapes

a new test problem class which can benefit efforts of rapid EA prototyping.

6.2 Return to Context

There are different types of resonator design tasks. Some are easy to solve like the sizing of an organ pipe or the layout of a radio receiver. Simple analytical equations can represent the behaviour. When constructing an acoustic resonator for sonoluminescence experiments, in the simplest case of a spherical thin-walled glass flask such an approach may still be taken. However, in order to understand the vibration behaviour of liquid-filled acoustic resonators with nontrivial geometries, it has been shown that simple analytical models are insufficient, yet detailed finite element models with fluid-structure interaction can fill part of the gap. In particular, the resonator design of West & Howlett which had been used by Taleyarkhan et al. at ORNL and Purdue for sonofusion experiments since 2002 falls into this category (other recently published SF resonator designs, e. g. the ones by Tessien et al. of Burst Laboratories [348, 399, 473] would fit just as well). FE model investigations of this resonator add a new perspective to the controversial debate sparked by Taleyarkhan's sonofusion experiments. It could very well be the case that while the radiation detection topic was discussed in depth the profane aspect of resonator mechanics was overshadowed and neglected while having a substantial influence on resonator performance and cavitation conditions in real-world laboratory setups.

The group around Taleyarkhan claims that past experiments had successful outcomes. The history of replication experiments by others is inconclusive. FE models of the resonator design show that its extremely high sensitivity to geometry dimensions offers an explanation for the observed wide spread of cavitation rates and the inconclusive outcomes of replication experiments which has not yet been seriously discussed in the topical literature before.

In this context the proposal of a better reproducible resonator design advances the debate. Suggesting well-performing and reproducible new resonator designs makes renewed experimental efforts for setting up and conducting Taleyarkhan-style SF experiments worthwhile again. Successful SF data has little value if the experimental setup is not reproducible by others.

Similarly, one single optimised resonator geometry is of much less value than a reliable resonator layout methodology. The present work describes a thorough process of resonator design optimisation with global search algorithms in the form of evolutionary algorithms. Global EA search and consecutive fine-tuning with a local downhill-simplex search allows one to get quickly from an initial design concept cast into a well-parametrised FE model to a thoroughly optimised setup exhibiting optimal or near-optimal vibration modes with maximised acoustic pressure amplitudes at the centre and minimised pressures near the walls. Only a good optimisation routine makes different design ideas comparable.

With a standardised and quick optimisation procedure in the toolbox, simulation work accompanying future experimental SF resonator investigation campaigns can become much more responsive than it has been in the past. As any experimental research work is a steady process of learning and refinement steps, and as simulations are crucial to gaining understanding about system behaviour, such a qualitative

change and a change in responsiveness on the side of numerical simulations can become decisive for the success of future experimental campaigns.

6.3 Outlook

At the end of this work of reviewing the status of SF, investigating the problem of sensitive resonator setups, and suggesting new designs and a design optimisation methodology, several options of further advancing the topic can be suggested:

- The newly proposed SF resonator geometries A & C can be built and investigated experimentally.
- Simulation models for accompanying experimental campaigns need to become more reliable, this can be achieved via model calibration based on carefully gathered benchmark data. A systematic stepwise approach of model calibration goes out from trivial systems like simple glass discs or pieces of piezoelectric ceramic to make a way forward over simple assembled geometries (e. g. piezo plate glued on glass disc) towards the final goal of high-fidelity models of entire liquid-filled acoustic resonators.
- The calibration of FEM simulations accompanying experiments can be used to measure material-dependent damping constants and to determine and fit loss models for the transducers.
- With an updated high-fidelity FE model at hand, local re-tuning of the proposed resonator designs will become necessary.
- With EAO runs it can be investigated whether an updated material library will have a substantial impact on the shape of the result designs.
- Resonator models can be investigated with modern approaches of response surface-based sensitivity analyses in order to highlight the robustness of properties locally around optimised design points. An efficient meta-model-based approach can capture and quantify nonlinear and coupled parameter influences [313].
- It might be interesting to investigate the question of "a posteriori tunability" of resonator designs, e. g. by grinding and thinning reserved geometry segments.
- Enabling the systematic tunability of resonators or assembly parts can make it easier to deal with unavoidable property differences between pairs of piezoceramic transducers in two-transducer setups.
- It might be worthwhile to investigate the trade-offs between systems with lower Q to be driven with stronger transducers versus aiming for high-Q systems and the related issues of system cooling.

Of course, the final goal is clear, it is the repetition of sonofusion trials with improved equipment and under the condition of reproducibility.

List of abbreviations

Abbreviation	Description
CMA-ES	evolution strategy with covariance matrix adaptation
EA	evolutionary algorithm
EAO	evolutionary algorithm optimisation (meaning optimisation by evolutionary algorithm)
EC	evolutionary computation
ES	evolution strategy
FE,FEM	finite element (method)
KIT	Karlsruhe Institute of Technology (Karlsruher Institut für Technologie)
NFL,NFLT	no free lunch (theorem)
ORNL	Oak Ridge National Laboratory
RPI	Rensselaer Polytechnic Institute
SF	sonofusion
SL	sonoluminescence
THEA	tier-based hybrid evolutionary algorithm

Appendices

Appendix A

Sonofusion in a nutshell

A.1 Sonoluminescence and the question of sonofusion

In order to understand what *sonofusion* means, it is necessary to understand the term *sonoluminescence (SL)*. Luminescence is nothing else than light emission and the prefix *sono* is derived from *sonus*, the Latin word for sound. Just as chemiluminescence describes light emission accompanying a chemical reaction (e.g. bioluminescence produced by fireflies) or radioluminescence stands for light created in connection with radioactivity, sonoluminescence was the word coined around eighty years ago describing small amounts of light emitted by liquids subject to heavy sound fields. In fact, sonoluminescence is the appearance of plasma as a consequence of compression heating inside collapsing cavitation bubbles, and the question of sonofusion is ultimately a question about the maximum temperatures and pressures that can be achieved with SL plasma.

A.1.1 Sound leading to cavitation

Sonoluminescence was discovered 80 years ago because at that time high-power sonar systems were newly available which triggered a lot of academic research on high-intensity sound waves. Sound waves are time- and space-dependent pressure modulation patterns travelling through a medium at the speed of sound. If the amplitude of a sound wave exceeds the ambient pressure, regions are created where the medium is under tension[1]. If the tension in a liquid is strong enough, it can rupture, and one says cavitation bubbles are created. Cavitation bubbles grow in the tension part of the acoustic cycle and shrink and implode a little later during the compression part of the cycle.

A.1.2 Cavitation leading to flashing plasma

What happens inside an imploding cavitation bubble? "Nothing much" could be the intuitive answer, considering that cavitation bubbles are not *pumped-up* bubbles like those in sparkling wine or boiling water. If sonic cavitation means rupturing liquid

[1]Not all types of fluid can be put under tension. States of tension are only possible in a liquid where the collective behaviour is determined by attractive inter-particle forces.

under tension, then these bubbles are *pulled-up* bubbles, they should be empty. So nothing much should happen upon closing the holes again, except perhaps some sound emission when liquid smashes against liquid. But reality behaves differently: there is no vacuum next to a liquid surface because evaporation and outgassing of dissolved gasses fill up the new empty volume. This stream into the gas phase is only stopped after the bubble has reached its maximum size and begins to shrink again, driven by the buildup of the acoustic pressure in the liquid outside. Now, if that shrinking process was slow enough, vapour would all condense and gasses, although a little more reluctantly than the vapours, would be driven back into solution. But as the collapse becomes very rapid (see figure A.1) near the end, there is just no time for the gasses to go back into solution, and this is when high pressure builds up inside the bubble as all the kinetic energy of the surrounding liquid plunging towards the bubble centre is loaded upon the spring of the bubble volume being compressed. Even if the bubble is only filled with more easily condensable vapours, this pressure buildup will happen for the last part of the collapse, when the condensation rate cannot keep up any more with the steep pressure rise. The compression leads also to a sharply rising temperature inside the bubble because the small time scale poses a hurdle to the heat transport as well. Today we know that it is this quick compression heating that leads to the bubble's content becoming so hot that it begins to glow. A plasma forms in the centre of the imploding bubble emitting the mostly blueish light of sonoluminescence.

A.1.3 Sonoluminescence plasma igniting the question of sonofusion

SL plasma doesn't glow dimly red, it doesn't glow white, it has a bright blueish shine. Is it thus hotter than the sun? Adding these three facts can make it plausible to expect extremely hot temperatures in SL plasma:

- The origin of SL is plasma lies indeed in inertial compression and heating (and not in e. g. the rupturing of the liquid similar to the cracking of a flint.[2]).
- Describing the bubble collapse with the equations of continuum mechanics while taking account of the finite speed of sound inside the bubble yields a singularity due to a concentric shock front developing in the gas phase and its spherical symmetry. Temperature, pressure, and density rocket towards infinity.
- From the side of experimental research on SL, since early on and throughout the years, a persistent fraction of publications has kept on suggesting "extreme temperatures".

If these theories and data suggest that breaching the fusion threshold could be possible, one may ask: what options are there to tweak a sonoluminescence experiment and tune up peak plasma temperatures to reach beyond 10^6 or 10^7 Kelvin? Adding some fusion-capable isotopes like deuterium or tritium to the liquid and gas cocktail, can thermonuclear fusion be provoked inside the little plasma flares? Next to

[2]More historic and current alternative explanations of SL are listed in chapter 1.3.2

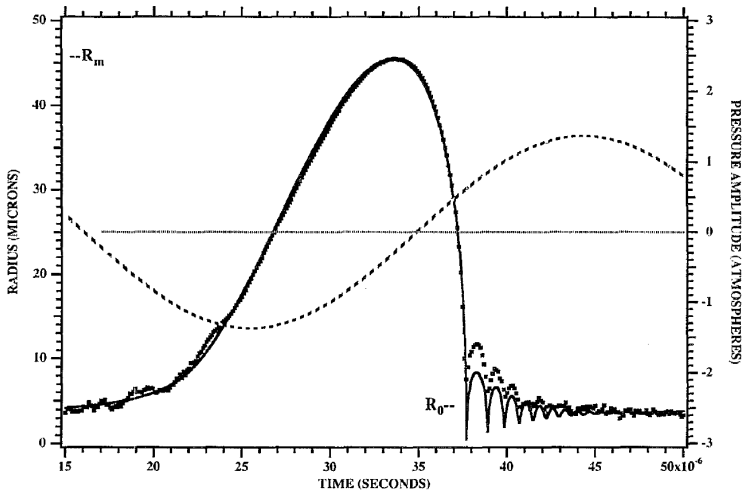


Figure A.1 Numerical simulation of the Rayleigh-Plesset equation by Löfstedt et al. [276].
The Rayleigh-Plesset (RP) equation (see chapter 1.3.1) describes the temporal development of the radius of a spherical cavitation bubble and is based on formulating the pressure equilibrium at the interface. The simulation by Löfstedt et al. accounts for the energy loss into the surrounding liquid through sound emission, but neglects the liquid's viscosity and surface tension. The gas volume was assumed to be of uniform density, pressure, and temperature obeying the thermodynamic laws of a hard-sphere Van-der-Waals gas. The dashed line is the sinusoidal acoustic driving pressure, the solid line is the simulated bubble radius, and the dots represent experimental bubble size data gained by light scattering. The amplitude of the acoustic driving pressure was 1.35 atm and the ambient pressure 1 atm. Therefore, the liquid surrounding the bubble is under tension between $t_1 = 21$ and $t_2 = 30$ microseconds. Bubble growth after t_2 is driven only by inertia of the liquid surrounding the bubble. The liquid's inertia in connection with the nonlinear spring characteristic of the hard-sphere gas and the radial geometry is responsible for the collapse speed showing no slow-down until very near the minimal radius and the sharp-edged-looking rebounce in this plot scale. (Reprinted from [276] with permission from S. Putterman and the publishing house.)

sonoluminescence, one would have *sonofusion (SF)*. Creating fusion-capable hot and dense plasma with the simple and low-tech experimental setup of SL (a liquid-filled glass flask, one or two piezo rings glued to it as vibration drivers) would indeed be something remarkable. Well, remarkable or rather completely irrealistic? There are also some thoughts suitable for curbing enthusiasm:

- Heat transfer through conduction and radiation are strong antagonists of energy concentration, particularly at the small scales of SL where the plasma's surface-to-volume ratio is very high.
- While heating the bubble content, if the chemical bonds of molecules need to be broken up, the resulting energy dissipation antagonises the energy concentration.
- The imploding bubble is quite a remarkable tool for energy concentration, for going from the very low energy densities of a sound field in liquid to the very high energy densities of incandescing gas. But extrapolating from currently available experimental facts, there is still a large gap to bridge to reach fusion conditions as can be seen from the following example: One of the hottest and densest experimental plasma condition estimations[3] comes from a spec-

[3]Many publications with more spectacular estimations may be found, but the data and interpre-

troscopic experiment performed by David J. Flannigan and Kenneth S. Suslick [133]. Using the formulae given in appendix B.1 one can infer that their sound field in 85 wt% sulphuric acid with an amplitude of 3.8 bar induced a strain energy of $2.5 \times 10^{-5}\,\mathrm{J/cm^3}$. Their central result is a lower bound for peak plasma conditions which they give in the form of an electron density $n_e = 4 \times 10^{21}\,\mathrm{cm^{-3}}$, the ion density $n_i = n_e/3$, and a temperature of $T = 16 \times 10^3\,\mathrm{K}$. This can be used to calculate a plasma energy density u of

$$u = \frac{U}{V} = \frac{3}{2} n_i k T_i + \frac{3}{2} n_e k T_e + \frac{4\sigma}{c} T_r^4 \approx 1.8 \times 10^3\,\mathrm{J/cm^3}$$

where k and σ are the Boltzmann and Stefan constants, and T_i, T_e, T_r are the ion, electron, and radiation temperatures, assumed all equal.

This means that the collapses of the oscillating bubble in that experiment induced an energy concentration over 8 orders of magnitude which is certainly remarkable, but it means that at least this experimental setup is still more than 6 orders of magnitude away from fusion conditions as e. g. achieved in laser fusion experiments at the US National Ignition Facility in Livermore, CA (see appendix B.2).

Whether fusion conditions can be achieved or not, this question cannot easily be answered from looking at the spectra of SL because the analysis might yield more optimistic or pessimistic answers depending on assumptions made on the origin of the light. One also cannot easily estimate answers based on calculations from first principle because many decisions between a multitude of different possible and reasonable model assumptions have to be made.

But maybe one should look at it from the other side: Why not take fusion as a measuring device for SL plasma conditions? If fusion does not happen, then the plasma is too cold and thin. If fusion happens and can be detected, well, then SL plasma must be getting dense and hot enough. In that case the measured fusion rates and pulse widths would allow us to learn more about the physics inside the imploding bubble and the data could be used for calibrating the model assumptions used in theoretical calculations.

Let's go back into the centre of the imploding bubble where sonoluminescence originates and examine in more detail the nature and ingredients of the phenomenon. Let us learn more about how a sonoluminescence experiment should look if it was aimed at high temperatures and towards sonofusion.

A.1.4 Plasma ≠ plasma

There are many sorts of plasma, ranging from the thin and cool plasma inside a neon tube to the hot and dense plasma of a hydrogen bomb. Where does SL fit in? What can the colour of SL, or better its light spectrum, teach us about the conditions inside a sonoluminescing bubble?

tation from Flannigan & Suslick has been chosen for this calculation because it was deemed much more reliable than other reports for reasons mentioned below in the more detailed description of the experiment in section A.1.5.

Hot pieces of charcoal and iron are two examples of a "black-body radiator". This means that from the colour of their glow one can reliably imply the temperature of the glowing surface because the emitted light spectrum is determined by Planck's Law. E. g. the embers in a campfire appear reddish at 700 °C, yellow at 1100 °C, and bright white at 1500 °C. The surface of the sun is another black-body of 6000 K, the one defining for us what is neutrally white. Stars cooler than the sun appear reddish, hotter stars blueish. SL being blue, does this mean it is hotter than 6000 K? Not necessarily, since among the different sorts of plasma only some are black-body radiators.

Most people have seen neon tubes and know that they are a sort of gas discharge lamp. Inside, electrons are forced to travel from one end to the other. If the tube was empty, the electrons could fly straight through like in old TV tubes. But gas discharge tubes contain a small amount of gas at low pressure. This means electrons can accelerate towards the anode for a while until they collide with the next atom. In these collisions valence electrons are excited or flipped away, turning the atom into a positively charged ion. As a consequence, throughout the tube de-excitation and recombination events (when an electron fills the valence band gap of an ion) take place, and in these events the energy that once came from the collision is given off again in the form of photons. As a result, the plasma of a neon tube is cool and its characteristic red colour comes from the neon atom's individual pattern of valence electron energy levels. A spectrometer reveals that the neon tube's light is composed of a couple of characteristic spectral lines representing the possible de-excitation transitions of the valence electrons. Discharge lamps with other gases have other spectral line patterns and produce other colours: orange for sodium, blue for krypton, ultra-violet (UV) for mercury[4]. The working principle of discharge lamps requires low pressures giving the electrons large enough mean free paths to accelerate. With the much shorter mean free paths at ambient pressure only much higher field strengths can maintain gas ionisation and the character of the phenomenon changes. Discharge plasma at ambient pressure is created in the form of arc discharge as in lightning or with a polyester jumper. Here, electron avalanches[5] are needed to create a conducting channel[6] of ionised gas which is then heated up. In the case of thunderstorm lightning, according to [423], the radius of the plasma channel is about 10 cm, a total charge of up to 25 Coulomb is transported while a potential difference up to 10^8 Volt is bridged from cloud to ground. The plasma inside the channel is heated up to 3×10^4 K, and around 50 % of the plasma's energy content is released in the form of electromagnetic radiation (the other half being heat, chemical reactions, and a tiny fraction going into sound). A bolt of lightning is indeed a black-body radiator, its light spectrum is not dominated by spectral lines. Its bright white shine with steely blueish colour is the result of it being four to five times hotter than the surface of the sun.

[4]Mercury is used for the white tubes for room lighting. The emitted UV light is transformed into visible light by the fluorescent inner coating of the glass tubes.

[5]For thunderstorm lightning, the exact mechanism seems still debated [37], avalanches of relativistic electrons initiated by cosmic ray air showers are gaining popularity in explaining lightning initiation [184, 356].

[6]The creation of the conductive channel consisting of a hose of ionised gas is called "breakdown" or "avalanche breakdown" of the insulating gas layer.

The two important points are:

- There are very different plasma regimes, thin, dense, relatively cold, extremely hot, emitting spectral lines or black-body radiation or else.
- Only the right theoretical assumptions allow a valid interpretation of measurement data.

For measuring and comparing the temperatures of a piece of coal glowing yellow, the sun, and a lightning bolt, knowing the assumption of black-body radiation is a good approximation in these cases, it is only necessary to compare the intensities of the different colours in the light spectrum. But estimating the temperature of the white neon lamp illuminating your office based on this assumption will fail. For that task you would need a different theory than the one behind black-body radiation, you would need exact quantitative models for what happens in the plasma and how the photon conversion in the coating works.

A.1.5 Back to SL plasma

Part of the difficulty of interpreting SL spectra arises because even within this tiny niche of plasma physics there are many types of setups producing very different results. Depending on the used liquids, dissolved gasses, preparation and purification methods, temperature regimes, bubble nucleation mechanisms and collapse dynamics plasmas of different sizes and densities will arise, and due to the quick heating rate it will not always be close to equilibrium. Then, for each type of experiment the data has to be treated differently in order to gain $T_{\max}$, $p_{\max}$, and $\varrho_{\max}$. It begins with the rather simple model assumptions of how any spectrum is distorted when it has to go through liquid and glass (you can just measure that with a calibrated light source), and what the hard part of the plasma radiation does to the cold liquid molecules surrounding the bubble (e. g. trigger additional light emission) and ends with rather subtle model assumptions of how much the plasma's outer regions are hiding the inner core, how far or close SL plasma is from thermodynamic equilibrium, or how far the effects of high electron density phenomena reach in diminishing the usefulness of large amounts of available theory intended for the interpretation of thinner plasma.

All the different ways of conducting a sonoluminescence experiment and all the modelling approaches and their manifold theory ingredients used to interpret the measured data make up a vast and vividly developing body of scientific literature. In the following paragraphs a few key conceptual elements for understanding the basics and the variety of sonoluminescence will be picked out and sometimes illustrated with interesting findings reported in literature. From the point of view of being interested in the question of sonofusion, obviously, all the factors playing a role in energy concentration are of special interest

Poking into SL plasma

SL plasma can produce both types of spectra depending on experimental setup, ones with sharp lines and others with continuous intensity over frequency. The ori-

gin of the *continuum* part of the spectrum has been the cause of two decades of debate (see chapter 1.3.2), and part of that debate involves the question whether the plasma core needs to be opaque and optically thick in order to explain particular observations. For one thing, the opaqueness is an indicator that certain lower bounds of temperature and density must have been exceeded. But it also means that the plasma core itself is hidden and can be hotter than what the observed spectra suggest. That imploding cavitation bubbles can indeed produce dense opaque plasma has been demonstrated only recently by Khalid et al. [232] in a very nice experiment. Instead of just looking at the little plasma ball they poked right into it: by shooting at it with a very strong laser and tracking the laser-plasma interaction. Their setup differs in a few aspects from usual SL experiments. Instead of an acoustic resonator they used a water hammer tube,[7] a little cylindrical, sealed glass vial with dehydrated phosphoric acid under 20 torr xenon inside. The vial is kept in upright position, vibrated vertically at 40 Hz with an amplitude of 1 mm and rotated around its own cylinder axis at 300 Hz. By laser pulse they seeded a bubble in the lower part of the liquid column that would settle in a stable central position 2 cm above the vial's bottom. Here it repeatedly implodes when the vial's movement changes from downwards to upwards while the liquid above the bubble still wants to keep on moving downwards. This process allows rather big bubbles (0.5 mm) to implode under substantially higher pressures than is usually the case in SL experiments based on sonic resonators, and it results in long luminescence flashes of around 500 ns duration (FWHM).

A candle flame, although one normally cannot look through it, is still transparent, as can be seen from the shadow of a burning candle. The transparency test by Khalid et al. is not about examining the plasma's shadow. The idea is instead to use a strong enough laser beam pointed on the plasma. If the plasma contains enough free electrons to become opaque then the high laser energy absorbed by the electrons would significantly change the plasma's energy balance, and this was made visible by comparing the laser-on and laser-off scenarios from the side view[8]. The experimenters' astonishing finding was that not only were the flashes from laser-treated plasma brighter and longer than the flashes from untouched plasma. The most interesting detail was that the plasma was heated asymmetrically by the laser, the plasma was so opaque that the laser could only penetrate the first part of the plasma cloud of about 100 µm in diameter. This allowed Khalid et al. to infer that

[7] A water hammer tube is a small slim glass cylinder, half filled with a liquid, then sealed while under vacuum, so there is little gas content – the volume where there is no liquid is mainly vapour. When the tube is being shaken the liquid plug moves back and forth in the vial. Each time it smashes against one end strong shock waves run through the liquid column and are transformed into tension waves upon reflection at the free surface. If the tube is kept upright and shaken vertically at millimetre amplitudes, then the liquid plug never smashes against the upper end of the vial, it just hits the lower end as long as there is a large bubble separating the liquid plug from the bottom end. Khalid et al. however report in [232] that there was no bubble below the examined bubble which was itself hovering 2 cm above the vial's bottom end. Therefore it must be assumed that it was not water hammer shock waves travelling through the liquid that forced the bubble into implosion, but rather that the top part of the liquid column above the bubble was kept smashing against the lower part with the bubble of varying size in between.

[8] The side view pictures were taken filtering out a different frequency band than the incoming laser covered, to exclude confusion with scattered light.

the mean free path of an electron in the plasma can at most be 85 µm. The authors also reported a control experiment: they tuned the laser to fire between 0.5 and 1.0 µs later, at times when the SL plasma had cooled down and only a dim glow remained. Thus they created a matrix of three experiment conditions with the two switches laser-on versus laser-off and the timing aiming at either early & bright or late & dim luminescence. It was revealed that equal power laser-heating on the dimly glowing plasma has a substantially smaller effect. This is in line with the expectation that cooler plasma with a lower concentration of free electrons should be more transparent and let most of the laser light pass right through without interaction.

The experiment is particularly instructive because it shows that besides just passively recording emitted light spectra and analysing the data with assumption-laden theories, one can also use the tool of laser-plasma interaction to check fundamental assumptions by actively poking the SL plasma and watching the reactions.

In the case of ultrasonic cavitation inside a traditional resonator there is just one report by Cao et al. [71] who tried laser-heating on an SL bubble in pure degassed water. They were, however, not able to observe any effect.

Traditional SL spectroscopy

There is no access with physical instruments (e. g. thermometer) to the inside of an imploding cavitation bubble. Instead, the possible information channels for learning about imploding bubbles and the conditions of SL plasma consist of the bubble size and shape (photography, light scattering at the bubble interface), the the sound created by the implosion (microphone), the newly produced chemicals accumulating in the liquid after long hours of sonoluminescence (i. e. the result of sonochemistry), and the analysis of the time and frequency properties of the emitted light, i. e. SL spectroscopy. The latter channel offers by far the most direct information about what happens inside the collapsing bubble. A look at two publication examples shall give a quick impression of the state of the art of interpreting SL spectra.

Hammer & Frommhold [186, 188] interpreted spectra from bubbles in water containing water vapour and noble gasses. They went out from the simple assumption of adiabatic compression of the bubble's interior, but composed a large and complicated simulation of what happens in the heated matter. It included evaporation and condensation (changing the number of particles inside the bubble), dissociation of chemical bonds, heat transfer to the surrounding liquid, ionisation (with corrections for high densities). The aim was to quantify the rates of many types of particle interactions in the plasma and to measure the production rates of photons from some of these interactions yieding the light spectrum and the energy lost by the plasma due to electromagnetic radiation. The quality measure was how well the calculated spectra fit the observed ones. While 1994 Frommhold & Atchley [148] still concentrated on a very special collision and radiation mechanism (so-called collision-induced emission, CIE) of plasmas to explain SL spectra and harvested criticism that their model contained "still too many indeterminate points and adjustable parameters to permit a judgement on its tenability" [121], Hammer & Frommhold stress in their paper of 2000 [188]: "No unphysical or arbitrarily chosen parameters are needed, if the model is combined with hydrodynamic stability calculations." Finally, one of the

most important later additions to the model described in [186] is to allow the gas bubbles' outer shell to become cooler and its core to become hotter, giving up the too simplistic assumption of homogeneous temperature distribution. That step took away from long-wavelength and added to short-wavelength photons and brought the calculated spectra again closer to reality. Whereas in the paper of 2000 they report good agreement only for bubbles containing heavier noble gasses assuming temperatures around 20 000 K, in 2002 they report good agreement (through more UV radiation) as well for SL from the lightest noble gas, helium, assuming a hot inner core (16 % in terms of diameter) is isotropically heated up to 59 000 K. But throughout the time, the model described the bubble content as a rather transparent, optically thin plasma, a volume emitter (no surface emitter like a blackbody), a gas where bremsstrahlung from collisions between electrons and atoms and the recombination of electrons with atoms are the dominant generation mechanisms for the light we see as SL. A strength of the model is the ability to correctly predict the overall photon output of one bubble implosion, and how it changes when the water temperature is being varied and more or less vapour end up in the bubble.

Two articles of 2005 and 2010 by Flannigan & Suslick [133, 134] stand for a different approach to SL interpretation. Instead of trying to calculate everything necessary to tell the story of a whole bubble oscillation cycle and compare the whole spectrum and overall photon output with measurements, they just looked at two particular and local features in the spectrum of argon atoms, and with the help of fundamental laws of thermodynamics and quantum mechanics they asked: what plasma conditions are necessary for a few spectral lines of argon to look like we see them from a sonoluminescing argon bubble? Suslick et al. concentrated a lot of their work on noble gas bubbles in highly concentrated aquæous solutions of sulphuric acid and also phosphoric acid or even their dehydrated forms because these liquids have two advantageous properties leading to bright SL over many bubble oscillation cycles: their low vapour pressure ensures few of these molecules enter the bubble where their chemical dissociation consumes energy, and the liquids have a relatively large tendency to take all kinds of chemical compounds left over by the plasma burns back into solution and keep the bubble content clean. Inside the SL plasma particle collisions lead to valence electrons being kicked into excited states or completely out of the atom's electron shell. When electrons fall back from excited states into lower energy states or the ground states they emit photons. The photons whose wavelength λ, frequency ν, and energy E are correlated via $E = \hbar\omega = h\nu$ and $\lambda = c/\nu$ carry an amount of energy corresponding to the difference between the states. The relative population of the electronic states along the energy scale tells about the statistics of collision energies and thus particle speeds occurring in the plasma, and that in turn allows inferring the temperature. On the other hand, the levels themselves of the electronic states on the energy scale undergo characteristic modifications in dense plasmas where one ion's field has effect on a close-by atom's electron orbitals. These effects can be calculated in the so-called second-order Stark theory and together with the particle energy statistics and electrons' state transition rates, and together with other effects (e. g. Doppler effect) this results in predicting small frequency shifts and a broadening of argon's spectral lines in a characteristic asymmetric manner. By quantifying relative heights of some spectral lines and

the asymmetry of others, estimations of temperature and density were deduced by Flannigan & Suslick [133] for the SL plasma. They conducted these estimations repeatedly while step-by-step increasing the sound pressure amplitude. Thus, for sound pressure amplitudes between 2.7 and 3.8 bar they deduced temperatures ranging from 7000 to 16 000 K and electron densities from 4×10^{17} to 4×10^{21} electrons per cubic centimetre. This maximal plasma electron density n_e, the authors remark referring to [386], was "comparable to that generated by the Lawrence Livermore National Laboratory Nova laser (1.8 kJ in 1 ns at 527 nm) in inertial confinement fusion experiments on a polyethylene target." In the older paper of 2005, reporting on similar SL experiments in sulphuric acid under argon, temperatures up to 15 000 K were inferred from the argon spectral lines. In that measurement spectral lines from oxygen were analysed indicating the existence of doubly ionised oxygen atoms. Assuming thermodynamic equilibrium this finding would be in contradiction with the temperature measurement from argon because taking the second electron away from an oxygen atom's electron shell is connected with an energy cost of 13 eV, and collisions bringing that large amount of energy are just very improbable to occur at 15 000 K.[9] Or in other words: to force oxygen atoms to populate a state 13 eV above the ground state in significant amounts, much larger temperatures are needed. Flannigan & Suslick describe their findings as hints towards the existence of a hot, dense, and opaque plasma core. It would mean the SL spectrum is mainly created in a cooler outer shell of the plasma cloud which hides an even hotter inner core, but high-energy particles from that core can induce higher-energy states in the outer shell, and also some light from deeper inside will overlay the light generated in the not completely intransparent outer areas. A last important fact apparent from the data in [134] is that while increasing the driving sound pressure still before the bubble stability limit is reached, the argon temperature determination method becomes impossible because the above-mentioned line broadening smears out the spectrum structure.

To sum up: the two examples described were chosen to exemplarily illustrate two general approaches of interpreting SL spectra. On the one hand there is the composition of ever more comprehensive and complex multi-scale and multi-phenomena models aiming for the ability to simulate everything of importance happening inside an oscillating luminescing bubble and distilling the correct overall spectra with time-dependence. On the other hand there is the interpretation of a particular feature of the spectrum based on calculations according to very fundamental laws of electrodynamics, thermodynamics, and quantum mechanics. The first approach brings a lot more understanding if it works but may often carry uncertainties and ambiguities because two model mistakes can compensate each other. The second approach yields much more limited statements, but with better reliability

In the early days of SL its existence raised big questions. The time-resolution of

[9] The ideal gas law can be written as $pV = NkT$, at the same time the total kinetic energy of N particles is $\frac{3}{2}NkT$. Hence, the quantity $\frac{3}{2}kT$ can be identified as the average particle kinetic energy $\overline{E_{\mathrm{kin}}}$. But in terms of thermodynamics the quantity $E_{\mathrm{th}} = kT$, called the thermal energy, is often more relevant. In plasma physics, instead of T in Kelvin often E_{th} is given in electronvolt. A thermal energy of 1 eV corresponds to a temperature of 11 605 K. Consequently, a temperature of 15 000 K implies an average kinetic energy of $\frac{15\,000}{11\,605} \cdot 1 \cdot \frac{3}{2} \approx 1.94$ electronvolt.

flash detection was so low so that it was impossible to tell whether the light is created during moments of bubble nucleation or collapse and whether the luminescence originates from compressing, colliding, or fracturing pieces of matter. Nowadays the view is much more settled that the hot and compressed bubble content turns to plasma and generates the light and the debate has shifted to discussing the relative weight of the different radiation mechanisms. But knowing, that it is not charge separation by friction or chemiluminescence (the light of fireflies) induced by radicals and broken molecules, knowing, that it is the simple accumulation of pure heat that gives rise to SL, the question of what mechanisms can explain such an extraordinary energy concentration becomes even more prominent.

A.1.6 The SL principle of imploding bubbles: intrinsic mechanisms of energy concentration

How does an imploding bubble collect widely spread potential energy and concentrate it into a small spot? The most simple story could look like this: imagine an almost empty cavitation bubble in water in an environment of rapidly rising pressure e. g. acoustic pressure from a sound field:

- The bubble begins to shrink, the liquid surrounding the bubble accelerates towards the centre, potential energy is taken from the sound field and turned into kinetic energy.
- Whatever remains inside the bubble, a gas like air or water vapour that has no time to condensate, gets compressed by the large volume of heavy liquid moving in the direction of an ever smaller bubble. The bubble content serves as a spring, taking up all the kinetic energy of the collapse and heating up in the process.

At the end, before the loaded spring starts to reverse the bubble contraction, all the available energy has been put into compression and heating of the bubble content, which momentarily covers only a tiny fraction of the bubble's initial volume.

Already from this simple thought experiment, some easy conclusion can be drawn about what conditions support energy concentration:

- The bubble must be large before it implodes, this will set more liquid mass in motion during collapse and faster interface speeds will be possible. It will lead to more kinetic energy being available for transfer into the bubble.
- The bubble should not desintegrate during the collapse. For this it has to stay in a parameter range where the surface tension keeping it together dominates over instability mechanisms leading away from spherical symmetry.
- The bubble needs to be as empty as possible. The final energy density will be higher if the available energy is distributed among fewer particles. The largest and brightest shining plasma might not be the hottest.
- The plasma can get hotter if less energy will be consumed with breaking up chemical bonds. This explains why bubbles filled with mainly noble gas atoms yield spectroscopy data implying very hot temperatures (see above, A.1.5).

- The liquid should have a low vapour pressure.
- Vapour should have a tendency to easily recondensate. This means the accomodation coefficient α should be high; $\alpha = n_{\text{absorbed}}/n_{\text{hit}}$ is the fraction of molecules or atoms that stay and enter the liquid phase after hitting the vapour-liquid interface from the vapour side.
- The liquid should have a low content of dissolved gasses. Gasses are much worse than vapours because during growth of a new bubble they are more likely to come out of solution and enter the bubble than they tend to return into solution during bubble contraction. During repetitive oscillations the bubble's gas content will tend to increase and pump up the bubble to a limit size. Such bubbles collapse with more cushioning than bubbles with only condensible vapour inside.
- The thermal conductivity of the gas inside the plasma should be low, heavy particles are desired under this respect.
- The gas particles should have few internal degrees of freedom (i. e. molecule rotation & vibration) to waste energy on. (The adiabatic index $\gamma = \frac{C_p}{C_V}$ should be high[10].)
- Another simple means of increasing the available collapse energy is to raise the ambient pressure outside the resonator. Assuming we have a liquid with a cavitation strength of 1 bar (that means it can bear tension up to $10^5\,\mathrm{N/m^2}$), then at one bar ambient pressure a sound pressure up to 2 bar can be realised. Raising however the ambient pressure to 10 bar will allow sound pressures up to 11 bar.[11]

More energy concentration mechanisms: concentric shock waves

Everybody knows: climbing slowly into the bathtub makes the water level rise evenly. Jumping in leads to temporarily very inhomogeneous water levels. This is not different for the density distribution inside the collapsing bubble. Up to now no assumptions have been made about how evenly or unevenly pressure, matter, and energy may be distributed within the bubble's internal volume during compression. In fact, if the collapse happens fast enough with relation to the internal sound velocity, density waves will be set in motion.

Water surface waves offer instructive examples of wave dynamics. When a stone falls into water it creates wave rings and the height of these waves decreases as they travel far away from the impact site. This is because the constant energy content of the wave structure has to be distributed along an ever increasing circumference. (The height decrease is not dominated by dissipation as would be the case in a liquid of higher viscosity like honey.) This implies that in the reverse case of concentric circular waves moving inwards the wave height has to grow because the circumference

[10] Under different considerations a low value of γ can become desirable, see [326] p. 3.

[11] For example a patent by R. D. Satterwhite [400] pursues a way of offsetting the pressure inside an acoustic resonator.

shrinks, and that it should theoretically go towards infinity when the waves hit the centre. That something in that direction is indeed the case can be seen in figure A.2 which shows a water column as the result of inward moving ring waves that were triggered to fill a dent in the water surface left by an impacting water droplet. In reality the height of the water column is not infinite because the surface tension antagonises the formation of needle-sharp structures and because the initial waves were broad and without steep slopes. The same mechanism is responsible when the energy of a tsunami wave is concentrated onto an ever smaller width in a V-shaped bay.

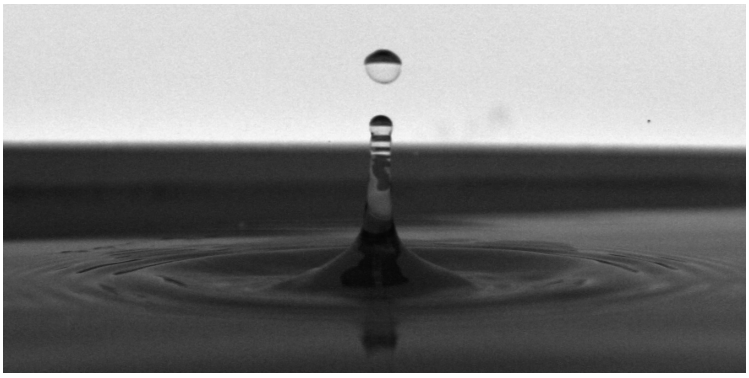


Figure A.2 Water column as a result of concentric inwards-moving waves.
When a stone or a water drop falls into water it displaces water and triggers concentric waves which move outwards. This outward movement of water results in a depression in the water surface. Gravitation then pulls neighbouring water back into the valley. This triggers a concentric wave travelling towards the centre of the depressed area. Just as all outwards moving ring waves decrease in height as their energy content has to be spread along an ever increasing circumference, the inwards moving wave grows in height as the circumference goes to zero.
This picture is not only useful for explaining energy concentration by concentric waves. It can also illustrate, that when there is a singularity (i. e. a denominator going towards zero, in the clean mathematical description) nature will almost always do something against it in the real world. So, there are several reasons why the water column cannot rise infinitely high. Most importantly in the concrete example, the water column cannot become smaller in diameter than the wavelength of the waves involved in composing the initial incoming ring wave, and if only a finite amount of energy is available from the initially falling water drop, then a finite-size water column can only be risen to a certain height with this energy. But even if the incoming wave was in the form of an almost vertical step, i. e. if it contained very short wavelengths, never could there be a sharp needle of water because of two antagonists that can be seen clearly in the image: surface tension and instability. Surface tension is responsible for the quick rounding of any arising sharp edges. Instability mechanisms with the help of the butterfly effect ensure that the thin water column never stays perfectly straight, symmetric, and evenly thick for very long. Instead, the water column wobbles into an amorphous but smooth shape and desintegrates into droplets. Hence, in the limit of very shallow depressions in the water surface with low water level gradients the large wavelength limits the energy concentration. In the case of deeper depressions with steep water walls clashing together at the centre the limits come from the surface tension and instability.

Turning back to the compression of vapours and gasses inside an imploding cavitation bubble, we can think of the rising pressure behaving similar to the rising tide in the bathtub. If the bubble narrows sufficiently slow then the pressure rises evenly throughout the bubble. But if the bubble wall moves fast enough then the pressure distribution will show wave-like inhomogeneities. The criterion here is the sound velocity in the gas phase, the velocity by which small-scale pressure differences are communicated across space. If the bubble wall approaches or surmounts this velocity, a density wave will accumulate in front of it and detach under certain

circumstances, leading to a scenario shown qualitatively in figure A.3.

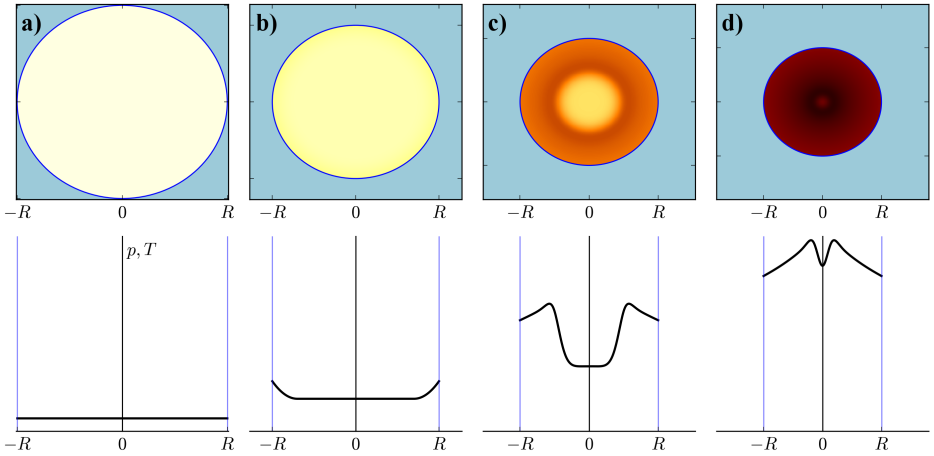


Figure A.3 Shock formation in the gas phase upon supersonic implosion.
Series of four snapshots during bubble implosion showing (only qualitatively without realistic proportions) a developing concentric supersonic shock front inside the bubble. a) Early stage of bubble implosion, the speed of the bubble interface v is still much slower than the speed of sound c inside the vapour phase, $v \ll c$. Pressure differences are communicated quasi-instantaneously throughout the bubble volume, thus the pressure distribution remains uniform. b) As v approaches c, the pressure distribution ceases to be uniform, a region of increased density is piling up in front of the bubble interface. c) Finally, the high-density pile-up turns into a shock, detaches, and runs ahead of the pushing liquid wall. This has less to do with changes in the two speeds of the sound and the liquid wall, and more with the ever increasing curvature and density difference. High curvature leads to the shock becoming quicker than the speed of sound. It turns into an increasingly supersonic shock. d) The macroscopic picture of the hydrodynamic shock running over the centre of the bubble translates into the microscopic picture of a moment when the gas particles in a region closely around the centre are pushed to fly towards it with high velocities and momenta. The result is inertially confined plasma flaring up for a short moment until the particles are repelled backwards, diluting and cooling the centre region again.

The shock wave inside the bubble can also be seen as the inverted version of the shock wave created by an detonation. The gasses created by exploding dynamite want to cover a large volume and the air has to be pushed away. This push comes in the form of a shock wave, a sudden jump in density and pressure forming a spherical wave front and travelling away from the detonation site at supersonic speeds. As it moves farther and farther away, the pressure and density differences across the shock reduce and so does the speed. The sharp pressure jump smoothens out and finally transforms into a sound wavelet which moves from that time onwards at the speed of sound. In the collapsing bubble it is just the other way round: the bubble interface, exceeding the speed of sound, begins to accumulate a density wave in front of it, and from now onwards, in principle no further increase of the interface speed is needed. The height of the density wave front increases by itself because of the spherical symmetry. Together with the rising curvature it entails a steady increase in speed responsible for the detachment of the shock wave from the bubble wall. In this scenario, with a concentric shock wave as a further mechanism of energy concentration, a fraction of the bubble's gas content receives an unproportionally high share of the available potential energy while the outer regions of the bubble stay cooler. Consequently, substantially higher maximal temperatures and pressures are reached in such a model compared to equating the bubble to an adiabatic piston.

What are the antagonists to this energy concentration mechanism, what prevents the temperature from diverging this time? Heat transfer and electromagnetic radiation of the plasma stay the principle drivers, but a different aspect comes into play as well: the failure of continuum mechanics in the microscopic scale. In a case where the equations of fluid mechanics let the shock wave steepen and rise only on the smallest scale, only on the last few nanometres towards the bubble centre, all the pressure, density, and temperature curves resulting from that math become meaningless because there cannot exist a collective movement pattern of smaller characteristic length than the particles' mean free path. If the equations say the temperature becomes infinite at $r = 0$ then the real world has no problem with it because it consists of particles with finite kinetic energies.

More energy concentration mechanisms: segregation of particle species

Numerical simulation of the collapsing bubble often means numerically solving a variant of the Rayleigh-Plesset equation for the bubble radius and solving the equations of continuum mechanics and plasma thermodynamics inside the bubble. But there is a different approach named "molecular dynamics (simulation)" (MD or MDS) which basically means the direct computation of the flight paths, collisions and interaction effects of an ensemble of many particles. If particles of different species with different weights are present inside the bubble, then the direct MD simulations are able to capture an important phenomenon that will be missed by a standard description in terms of continuum mechanics. Since at the same temperature light particles have much higher velocities than heavier ones, they can leave any area of locally increased density and temperature more quickly than the heavier species. In an article by A. Bass et al. [30] the result of a simulation with helium and xenon atoms can be seen. The effect allows the lighter helium atoms to diffuse out of a forming shock front. The shock front is then mainly formed by the heavier xenon atoms, and one could say a density peak of helium atoms "surfs" in front of the xenon wave. When this combined shock front runs over the centre, a population of helium atoms gets compressed by a wall of heavier xenon atoms coming up from behind. This is the same energy concentration mechanism as the moving heavy liquid compressing the bubble. In [29] the same authors present an MDS with pure xenon reaching a maximum temperature larger than 1×10^6 K for 0.1 ps (fig. 4 on p. 2126), but with the addition of 10 % helium the simulation yields $T > 1 \times 10^7$ K for 0.25 ps ([30] p. 4).

More energy concentration mechanisms: bubble cloud dynamics

What happens if there is not only one single cavitation bubble, but a whole cloud of them? The above descriptions often addressed a single bubble oscillating and flashing in a sound field (single-bubble sonoluminescence, SBSL) and sometimes mentioned *freshly nucleated bubbles* filled with mainly vapour and almost no noncondensible gas. Bubbles of the second sort, owing to the nature of bubble nucleation mechanisms (see appendix H), usually come in large numbers. Luminescence in this case is called multi-bubble-luminescence (MBSL). Also in the sonofusion experiments of

Taleyarkhan et al., where cavitation in acetone under tension is nucleated with neutron scattering, clouds of cavitation bubbles suddenly appear and go through a few cycles of expansion and collapse before disappearing again. A few thoughts should be spent on how MBSL differs from SBSL.

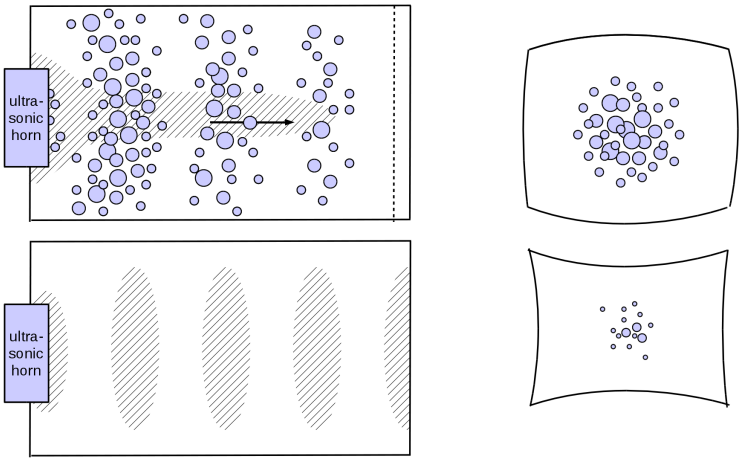


Figure A.4 Cavitation in front of an ultrasonic horn (left) versus cavitation inside an acoustic resonator (right). In the upper left image a source of ultrasound creates pressure waves that travel through the chamber filled with liquid from left to right where they are absorbed. For a human watching such a basin the whole volume will appear fogged because our eyes cannot resolve the quick succession of waves of bubble growth followed by fronts of bubble implosion. The sketch symbolises a snapshot where bubbly regions indicate where the sound wave generates cavitating low-pressure zones. Cavitation has the tendency to equilibrate high and low pressures by bubble expansion and contraction. Therefore, the amplitude is weakened as the sound wave travels through cavitating liquid. The hatched area shows a possible region of luminescence. The largest sound pressure amplitude occurs obviously directly at the surface of the sound emitter. The cavitation strength at this surface limits the driving amplitude. The sketch below shows an inefficient resonator by simply introducing the reflecting wall on the right. Here, standing waves can be excited allowing the sound pressures in the liquid volume to exceed the sound pressure on the sound emitting surface by a margin. Contrasting the travelling sound wave on the upper left, the drawing on the right symbolises a generic (e. g. cylindrical) good resonator creating a standing sound wave with the following properties: there is one single pressure antinode on the centre and it corresponds to a displacement node. The maximum sound pressure in the liquid can be much greater than anywhere on a surface. When the hull of the resonator expands and the liquid is under tension this represents a good moment for seeding cavitation bubbles which will grow quickly. Later during the acoustic cycle when the pressure rises throughout the volume, the bubble cloud will shield the centre area from the pressure rise. Since the liquid surrounding the bubbles has inertia, it takes a while until bubble growth is reversed to bubble contraction. The high-pressure signal advances slowly through the bubble cloud because only after the current outer layer of bubbles has collapsed completely the pressure rise can be communicated to the next layer inwards. A pressure sensor positioned at the outer rim of the bubble cloud would record the steady rise of the sinusoidal acoustic signal, whereas a pressure sensor placed further inwards would stay at vapour pressure for a while until at a later moment the pressure rise arrives with a much steeper slope in time.

The most important rule is: if there are cavitation bubbles nearby, then the liquid cannot be under tension nor can there be high pressure. Any tension or pressure in an area with many bubbles can immediately be equilibrated by a bit of bubble growth or shrinkage. As a sound wave in liquid consists of nothing else than elastic strain and compression, cavitating liquid dissipates sound. This is symbolically shown in figure A.4 (left) where a sound source on the left emits sound waves moving to the right and where the decreasing concentration of bubbles in the low-pressure areas from left to right indicates a fast decay of the sound amplitude. The hatched area in that drawing indicates a possible area of SL. The drawing symbolises a setup of

waves created by the ultrasonic horn on the left and travelling to the right where they leave the area of interest or are absorbed. In this setup the largest sound pressure amplitude occurs directly at the surface of the sound emitter. Cavitation on that surface and cavitation damping the sound wave throughout the volume limit the achievable maximum sound pressures. In the usual procedure associated with this type of experiment, cavitation bubbles are nucleated by impurities or pre-existing microbubbles present throughout the liquid. Because of the limited driving amplitude and because the cavitation bubbles in such a setup are of tiny micrometre sizes and cannot oscillate in their own resonance frequencies (which leads to small ratios of extremal bubble radii $R_{\max}/R_{\min}$), SL plasma in this case is cooler, around 3000 to 5000 K and under pressures of about 10^3 bar [284], and the emitted light spectra consist almost entirely of atomic and molecular emission lines and have a low continuum content. This kind of experimental setup seems to have led to the wide-spread notion in SL literature that MBSL creates *the less hot version of SL plasma.*

But inside a good resonator and employing the right nucleation technique a different regime of MBSL is possible. A good resonator is able to induce very high sound pressures deep inside the liquid volume while keeping sound pressures low where the liquid is in contact with solid materials. This is the only way to put liquid under large tension while preventing cavitation on solid surfaces from negatively affecting the sound field. Liquid impurities or microbubbles are harmful in this case and would prevent the buildup of large tension. Aiming for the largest possible bubbles, bubble nucleation is only desired at the moment of largest tension. To introduce a nucleation site right at the centre of the liquid volume and only at the desired time is only possible with radiation-based methods like laser pulses or by neutron scattering. Neutron scattering in acetone has been used in the sonofusion experiment by Taleyarkhan et al. and it doesn't create just one bubble, but a whole cluster of many bubbles. This can be beneficial for the maximum SL plasma temperature because of the way how a bubble cloud reacts to the sound pressure when it turns positive.

Deep inside the bubble cluster the pressure is kept at the liquid's vapour pressure by the surrounding bubbles. If the pressure rises slightly, vapour will condensate and the bubbles contract. If it sinks slightly, bubbles will expand. This means the bubble cluster serves as a shielding against the acoustic pressure which can not directly reach into it. When the sinusoidal pressure produced by the resonator around the cluster turns from strain to compression, layer after layer of bubbles has to collapse completely before the pressure rise can be measured in the centre position. This takes its time due to the inertia of the liquid surrounding the bubbles. Only after one layer has collapsed, the next layer will become subject to the pressure rise and the bubble collapse processes begin with more and more delay. For each layer the pressure rise does not only occur later but also with shorter rise time. The cluster bubbles collapsing layer by layer from the outer rim inwards average out into another concentric density wave with a similar energy focusing effect as the ones discussed above. That the pressure rise for the last few bubbles in the middle of the former cluster does not only come delayed and with a shorter rise time, but that its amplitude is multiplied immensely as well, can be seen from the calculations by

Nigmatulin et al. [326] (figure 6).

In order to summarise the three secondary energy concentration mechanisms, shock, species segregation, and cluster dynamics:

- An imploding spherical shock front can form inside a collapsing bubble, but only if the bubble implodes fast enough and the speed of sound inside the bubble is sufficiently slow.
- If it occurs, it represents another very similar mechanism of energy concentration on top of the bubble implosion itself.
- If there are two particle species present, it is possible that the lighter atoms begin to surf in front of the shock wave of the heavier atoms. The light particles will exhibit substantially increased peak temperatures when being smashed in the centre by the shock front of heavier ones.
- If a cavitation bubble cluster of spherical shape implodes under the influence of a rising external pressure, then that pressure signal reaches the centre of the cluster delayed but with increased steepness because the outer bubbles serve as a shield against the pressure rise until they implode. The progressive collective bubble collapse from the outer rim of the cluster inwards represents yet another spherical concentric wave which can amplify the external pressure signal.

A.1.7 Can SL plasma generate fusion?

Wax burns when it is evaporated to form a gas, when it is intermixed with oxygen, when that gas mixture is heated so much that its constituents, the paraffin and oxygen molecules break into pieces upon collision, and when the pieces (hydrogen, carbon, oxygen atoms and radicals, CO, C-clusters etc.) recombine into H_2O and CO_2 at much lower energy levels. The whole process is an exothermal reaction because the energy to speed up, collide, and break apart the input molecules is exceeded by the energy freed upon recombination into the products. The energy difference goes into the formation of photons, the light emitted by a flame, and into kinetic energy of the newly-formed molecules which is the heat produced by the flame.

Fusing light atomic nuclei into heavier ones is a reaction with very similar characteristics. When a star forms in space, gravitation pulls together cold clouds of mainly hydrogen. While for chemical reactions molecules need to be broken apart, for fusion it is the atoms that need to unravel. A contracting hydrogen cloud only becomes a star, the fusion chain reaction at the centre only gets started, under the condition that this environment is heated enough that atomic nuclei are stripped of their electron hulls (i. e. plasma is formed) and that the nuclei themselves have enough kinetic energy, that their collisions become violent enough so that a nonvanishing probability for various fusion reactions exists among them. Analogously to exothermal chemical reactions where molecules are broken up and regrouped into products of lower potential energy, fusing light nuclei to heavier ones is an exothermal reaction of breaking up and regrouping protons and neutrons into new combinations

of lower potential energy. Once the fusion chain reaction has been started at the centre of a new star, it contributes to the heating which is a positive feedback on the rate of violent collisions that lead to fusion reactions. A star can only exist if it has enough mass so that its gravitation sufficiently counteracts the increased pressure from the heat inside.

The temperatures necessary for fusion reactions ($> 10^7$ K) exceed by far what is normally possible on earth, but the question whether a collapsing cavitation bubble can ignite a "star in a jar"[12] keeps being raised in scientific literature because the maths describing the spherical implosion processes, due to a singularity, in principle allow it. The real question is just how far a real-world bubble will allow the comparison with the mathematical counterpart, meaning, how long before an antagonising mechanism like shape instability[13] stops further energy concentration leaving it up to heat transfer and radiation to dilute it again. On the other hand, it has to be kept in mind that there are no sharp fusion threshold[14] lines along the axes for temperature and density, there are only microscopic reaction probabilities and macroscopic reaction rates slowly rising above zero. This means that a little below the nanometre scale it does not matter any more if a shock wave were still concentric or not or whether there is any at all because the remaining nuclei within such small radii are so few that they cannot contribute a lot of traceable fusion products anyway – if the hottest central fraction of the plasma volume is too small, it may still allow fusion reactions in theory, but in reality the number of actually happening reactions may drop into the range of negligibility.

The interplay of theoretical predictions and experiments in SL and SF could perhaps be seen in the following way: any advance in theoretical modelling yields further understanding and allows for tweaking experiments to more extreme plasma conditions, and any advance in experimental techniques and any recorded SL spectrum with yet unseen characteristics brings further calibration input allowing improvements on the theoretical modelling side. Fusion neutrons, should they once emanate plentifully from an SL experiment, would be a set of calibration data of valuable information content, exactly because of how continuously the reaction rates scale with plasma density and temperature, and because of how directly this can be deduced from fundamental particle physics. Neutrons have no charge and therefore fly through solid matter over macroscopic distances rather undisturbed, unlike other particles, and the little disturbance they get on the way from the plasma to the detector can easily be recalculated from fundamental physics. This makes the neutron signal a very clean signal. The fusion neutron flux would be a high-quality

[12] The term has been coined in [369].

[13] Just like the water column in figure A.2 can't rise that high if it becomes deformed asymmetrically, an imploding bubble is hindered in its energy concentration if it becomes deformed.

[14] Many chemical reactions require a certain amount of activation energy, i. e. an energy barrier which has to be surmounted before the system can start to roll down the energy slope to reach the state of lowered energy after an exothermal reaction. However, in nuclear reactions the tunnel effect plays a substantial role. Two nuclei which are always positively charged and repel each other do not always have to surmount the whole energy barrier given by the repulsive potential. Due to the tunnel effect they can tunnel through the last bit of the barrier below the peak energy level. This is why there is no sharp energy threshold for nuclear fusion reactions, and hence, there are no sharp temperature and pressure limits.

integral measure telling about SL plasma quality – and it already is: measuring no significant neutron flux can give valuable upper-limit information.

So, can SL plasma generate fusion? Reasonable answers are: theory, on a fundamental level says "why not?", experiments at the moment suggest "rather not"[15], and a general look at SL with its manifold appearances and multi-scale nature leads to the thought that "there are many screws to tune".

A.2 A sober argumentation for sonofusion – Motivating sonofusion experiments without hyping them

Would sonofusion solve humanity's energy problem, as suggested in e. g. [66, 346, 383]? Let's clean up with the misconception traps induced by catchy press phrases starting along the line of *harnessing fusion on the tabletop* and ending in the extreme case with *saving the world with energy from sonofusion reactors.*

- **Harnessing fusion:** That term has different meanings in big science (laser-confinement fusion (LCF) and magnetic-confinement fusion (MCF)) and small-scale fundamental research. In the first framework "harnessing" means much more: proving that it is possible to run large quantities fusion reactions on a daily schedule with a long-term stable machinery and demonstrate economical electricity production from that machinery. In the latter case, for a team of researchers conducting SF experiments or other fundamental research like [160, 320] the main interest is this: directly tracing back some significant number of fusion reactions. Here, "achieving" or "enabling fusion" are correct translations. Sometimes, the term "harnessing" can also be interpreted as making the distinction between a *controlled fusion chain reaction in a laboratory or technical environment* and the *uncontrolled* version, the hydrogen bomb.

- **Would tabletop fusion be anything new?** No. Portable accelerator tubes triggering D-D or D-T fusion reactions to be used as neutron source are commercially available. They are used for imaging in the oil industry, archaeology, and in other areas. A little linear accelerator shoots deuterium (D) or tritium (T) ions towards D or T atoms embedded in a solid target. Ever smaller neutron sources based on this principle are being built. Net energy gain and sustained fusion chain reactions are not feasible with accelerated particle beams bombarding solid targets. Pyro fusion [320] is another example of tabletop fusion in research.

- **Would *thermal* tabletop fusion be anything new?** Yes. It must be noted that all the currently feasible ways enabling tabletop fusion supply the initial energy for overcoming the energetic barriers for fusion reactions with the help of the principle of particle accelerators. Thermal tabletop fusion

[15] If the status of the mainstream scientific literature may be summed up in two words, it may be this formulation of incomplete negation. Since the SF trials by Taleyarkhan et al. who claimed successful outcomes, the scientific community has staged replication trials only at a low rate. As argued in chapter 1.4, the status of reproducibility is inconclusive to this date.

where the energy for sufficiently violent particle collisions is supplied by the brownian movement of particles would indeed be something new. It would be quite noteworthy scientific news to know for sure that the thinly distributed energy of a sound field can be focused by a simple self-organising mechanism – spherically imploding bubbles – to heat a small volume up to tens of millions of degrees Kelvin.

- **Does a successful sonofusion experiment imply a controlled chain reaction?** No. There is a big difference between producing a few fusion reactions which can be detected and those high fusion rates that have a thermal feedback on the driving machinery or even plasma conditions. A sustained fusion chain reaction means that the plasma has to be kept stable for a while so the enthalpy of many fusion reactions substantially contributes to compensating the plasma's heat losses and maintaining conditions for many subsequent generations of fusion reactions. There are actually two levels of sustainability to talk about. In the context of a star, a hydrogen bomb or a magnetic confinement fusion reactor the fusion enthalpy has to compensate the energy losses of the plasma and prevent it from cooling down. In the context of imagining energy harvesting by SF (or other pulsed fusion machinery approaches) the fusion enthalpy would be needed to sustain only the pulsating machinery, not the plasma conditions. Currently, SF experiments are merely about detecting the phenomenon. Closing the energy cycle, i. e. using energy of SF events to replace part or all of the energy necessary for triggering subsequent cavitation and SF events is still science fiction.

- **Does a successful sonofusion experiment prove that electricity production is possible that way?** No. Sonofusion experiments are relying on the collapse of almost empty bubbles. This requires a low vapour pressure. With water-based liquids as they are used mostly for SL or with conventional organic solvents as used in the sonofusion experiments of Taleyarkhan et al. [458] low vapour pressure is ensured by low temperatures. Any energy from fusion reactions is in the form of kinetic energy of the fusion products, kinetic energy which is immediately deposited and turned into heat in the surrounding materials. Therefore, in practice, the energy of nuclear reactions can only be used to produce electric power with a conventional thermodynamic process like a steam cycle with turbine exploiting a temperature gradient and transferring heat from hot to cold. As long as the energy of sonofusion reactions is dumped into a liquid that has to be kept at lower than ambient temperatures, there is no prospect for energy harvesting. If one were able to demonstrate sonofusion in a liquid at hot temperatures, say, 500 °C (e. g. in a liquid metal), then the story would get more interesting. Besides the change in temperature regime, there is the change in scale to overcome. One would have to demonstrate SF in scalable irregular multibubble environments or in more complex sound fields with many pressure antinodes or in stacks of small resonators able to produce many sonofusion events in parallel.

Otto Hahn demonstrated nuclear fission on a tabletop[16]. From there, a long way of inventive engineering and upscaling led to nowadays' fission power plants which are obviously no tabletop equipment any more. If any of today's tabletop sonofusion experiments with cold liquids proved to be successful and reproducible, it would not tell anything about (a) a possibility to modify the experiment to work at high temperatures, and (b) there would of course be no guarantee that a likewise feasible upscaling route exists. But if reliable cool-temperature SF became demonstratable, it could be assumed that it would substantially trigger and increase research efforts aiming for shifting the SF regime up towards ever hotter liquid temperatures, and only thereafter, the issues of scalability and heat extraction from the liquid will become really pressing.

Imploding bubbles is a remarkably simple tool to create hot and dense plasma. Sonoluminescence provides an open door to examining matter at high energy densities with technology affordable for any university lab. Together with other means of examining plasmas, e. g. implosion of small magnetised plasmas [155] (medium-scale research) or facilities like NIF or ITER (large-scale research), comprehensive data can be generated. Each technique allows for different types of instrumentation and creates plasmas in different regimes. Combining and analysing new incoming data from the different techniques is of course highly beneficial for advancing the theoretical description of plasma physics. And as the theoretical understanding deepens, conclusions can be drawn on how to improve experiment designs and push the limits of feasibility ever further and on how to distil ever cleaner and more reliable information from recorded raw data. If SL can be pushed over the fusion threshold, a new and very clean type of data would become available for characterising SL plasma. In the case of opaque plasma, the in-depth neutron data would be a valuable complement to the surface light emission.

SL is remarkable for its energy concentration power, no microwave heating is needed to create the hot plasma, no femtosecond laser pulse, no explosives, no garage full of capacitors to be discharged in an instant, only a modest sound field suffices. Understanding the many facets of this multi-scale phenomenon has remained a hard problem for researchers for eight decades. The question of sonofusion has been one ingredient in the mix which is catching the fundamental researcher's curiosity. The answer to the question whether it is worthwhile looking at SF with the hope for energy harvesting will eventually also come as a result of this curiosity.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
C_p, C_V	heat capacity at constant pressure/volume
c	speed of sound/light
E	energy

[16] A reconstruction with original parts of the working table used by Otto Hahn, Lise Meitner and Fritz Straßmann for the discovery of nuclear fission in 1938 is at display at the Deutsches Museum in Munich.

$h, \hbar$	Planck constant ("quantum of action")
k, k_B	Boltzmann constant
N	particle number
n, n_e, n_i	particle/electron/ion density
p	pressure
R	bubble radius
T	temperature
T_e, T_i, T_r	temperature of the electron/ion/photon population
U	internal energy
u	internal energy per volume, e. g. plasma energy density
u_e, u_i, u_r	internal energy per volume of electron/ion/photon population
V	volume
v	velocity

List of Greek quantity symbols

Symbol	**Description**
α	accomodation coefficient
γ	adiabatic index or heat capacity ratio
λ	wavelength
ν	frequency, e. g. of electromagnetic radiation
ϱ	density
σ	Stefan constant
ω	frequency, e. g. of electromagnetic radiation

List of abbreviations

Abbreviation	**Description**
CIE	collision-induced emission
FWHM	full width at half maximum
ITER	International Thermonuclear Experimental Reactor
LCF	laser confinement fusion
MBSL	multi-bubble sonoluminescence
MCF	magnetic confinement fusion
MD,MDS	molecular dynamics (simulation)
NIF	National Ignition Facility
RP	Rayleigh-Plesset
SBSL	single-bubble sonoluminescence
SF	sonofusion
SL	sonoluminescence

Appendix B

Energy densities of sound fields and plasmas

B.1 Energy densities of sound fields

A remarkable feature of SL is the degree of energy concentration going out from the energy density level of the surrounding sound field and reaching up to the level of the plasma inside the collapsing bubble. Therefore it is interesting to put the energy densities of sound fields and plasmas in relation. From intuition based on everyday experience it is already clear that audible sound in air can be caused by quite small amounts of energy – it costs practically no effort to say a word or ping a glass – compared to the amounts of energy it costs to create heat – according to Wikipedia [44] it takes four healthy men or two well-trained athletes on bicycles to sustain the 1 kW power input to a stove plate.

The energy density of a sound field can be inferred by considering the energy density of elastic matter under compression or tension. This can be calculated using the generalised Hooke's law. The one-dimensional law $F = -kx$ can be turned into a corresponding equation for the three-dimensional continuum-mechanical case

$$\Delta p = -K\frac{\Delta V}{V} \tag{B.1}$$

where K is the bulk modulus and Δp the change in force per surface necessary to achieve a volume change ΔV. The potential energy U per volume V of a compressed elastic material is the work needed to bring it into the compressed state, and it can be gained by integrating Hooke's law, yielding

$$u = \frac{U}{V} = \frac{1}{2}K\left(\frac{\Delta V}{V}\right)^2. \tag{B.2}$$

Using once again eq. B.1 and the equivalence $K = \varrho c^2$ this can be further simplified:

$$u = \frac{U}{V} = \frac{1}{2}K\left(\frac{\Delta V}{V}\right)^2 = \frac{1}{2}K\left(\frac{\Delta p}{K}\right)^2 = \frac{p^2}{2\varrho c^2}. \tag{B.3}$$

This equation can be used to compile the energy densities of a few cases of interest given in table B.1.

Table B.1 Energy densities of acoustic pressure fields
The listed data contains energy densities for acoustic fields in air, water, sulphuric acid, and acetone. In the case of air there are three data sets corresponding to sound pressure levels of −6, 0, and 100 dB. 0 dB is the conventionally agreed-on hearing threshold. The lower value of −6 dB is an empirical hearing threshold of young healthy humans determined by Kurakata & Mizunami [245] (fig. 3). 100 dB corresponds to very loud noise (e. g. jet engine, discotheque) which can cause ear damage after extended exposure. For the case of water the acoustic pressure amplitude of 1.2 bar was chosen because it is indicated by Cheeke [85] as the onset of single-bubble sonoluminescence in pure water. The values given for sulphuric acid correspond to the strongest driving point of the multi-bubble sonoluminescence setup by Flannigan & Suslick [133]. Finally, in the last row conditions achievable in acetone are given. The sound pressure of 14 bar is twice the threshold of neutron-induced cavitation bubble nucleation, as required by the SF protocol of Taleyarkhan et al..

description	L_p [dB]	p [Pa]	ϱ [g/cm^3]	c [m/s]	u [J/cm^3]
hearing thresh. (young)	-6	1.0×10^{-5}	1.2×10^{-3}	343.2	3.5×10^{-22}
hearing thresh. (0 dB)	0	2.0×10^{-5}	”	”	1.4×10^{-21}
air 100 dB	100	2	”	”	1.4×10^{-11}
SL conditions in water	196	1.2×10^{5}	1	1482	3.3×10^{-6}
SL in sulphuric acid	206	3.8×10^{5}	1.84	1257.6	2.5×10^{-5}
acetone	217	14×10^{5}	0.791	1061.5	1.1×10^{-3}

The equations for the conversion between the sound pressure level L_p in decibel and the sound pressure amplitude in Pascal used for the compilation of the values in table B.1 are [262]

$$p = p_0 \cdot 10^{\frac{L_p}{20}} \qquad \text{and} \qquad L_p = 20 \log_{10}\left(\frac{p}{p_0}\right) \quad \text{with} \quad p_0 = 20\,\mu\text{Pa}. \tag{B.4}$$

The logarithmic definition of the sound pressure level serves the convenience of transforming a range of many orders of magnitude into a more easily memorisable number range. The transformation creates the necessity of an absolute reference in terms of pressure. Commonly, this reference pressure p_0 is set to 20 µPa which is an approximative value of the hearing threshold around 1 kHz [262].

An interesting piece of information one can take from table B.1 is that purified acetone through its robustness against cavitation allows a level of sound field energy density which is not really that tiny any more. 1.1×10^{-3} Joule per cubic centimetre is the same energy amount as can be stored by lifting the cubic centimetre of acetone up 14 cm. 1.1×10^{-3} J/cm^3 is almost two orders of magnitude more that in the case of the setup by Flannigan & Suslick who offer one of the highest reliable SL temperature deductions from experimental data in recent literature, and it's two and a half orders of magnitude more than reachable with popular SL setups in water. Water has no great cavitation strength because it does not easily release its residual dissolved gas content. That such a high energy content has to be built up in an acoustic resonator filled with acetone also underlines the main function of a good resonator: it has to couple the structure with the sound field and it has to allow the steady accumulation of kinetic and potential energy in the fluid over many oscillation cycles without damping the energy away.

B.2 Energy densities of exemplary plasmas

Now let us compare these acoustic energy densities with sonoluminescence and other plasmas. We can calculate the energy densities of a few sample cases of plasmas via the following equation

$$u = \frac{U}{V} = u_i + u_e + u_r = \frac{3}{2} n_i k T_i + \frac{3}{2} n_e k T_e + \frac{4\sigma}{c} T_r^4 \tag{B.5}$$

which sums over the energy contents of the ensembles of ions, electrons and photons (radiation). In the case of SL in sulphuric acid reported by Flannigan & Suslick [133] the article indicates the electron density $n_e = 4 \times 10^{21}\,\mathrm{cm}^{-3}$ directly and allows to infer the ion density from the degree of ionisation $Z = 3$ as $n_i = n_e/3$. Given a temperature of 16 000 K, equation B.5 evaluates to $u = 1.8 \times 10^3\,\mathrm{J\,cm}^{-3}$, assuming $T_e = T_i$. On the one hand the value can be seen as an approximative over-estimation because in the highly transient SL plasma the temperature of the electron system T_e may lag behind T_i [326], on the other hand Flannigan & Suslick point out that their SL experiment may have reached a regime where the surface layer of a sphere of dense and opaque SL plasma partially shields the radiation from an even hotter inner core. The comparison with table B.1 shows that from the acoustic pressure field with $u = 2.5 \times 10^{-5}\,\mathrm{J\,cm}^{-3}$ the energy density was raised by the energy focusing mechanism of the oscillating SBSL bubble by close to eight orders of magnitude.

The most well-known fusion reactor is the sun. The conditions in its core are assumed to be a temperature of 1.6×10^7 K and a density of $1.6 \times 10^5\,\mathrm{kg/m}^3$ [179]. The density implies $n = 9.8 \times 10^{25}$ nucleons per cubic centimetre. According to Djorgovski [116] and Serenelli et al. [412] the fractions of hydrogen and helium are 34 and 64 % with 2 % of other elements. Neglecting the heavier elements the particle densities for hydrogen and helium can be inferred via $4n_{\mathrm{He}} + n_{\mathrm{H}} = n$ as $n_{\mathrm{H}} = 1.1 \times 10^{25}$ and $n_{\mathrm{He}} = 2.2 \times 10^{25}$. With that, the energy density values can be calculated via equation B.5 yielding $u = 2.9 \times 10^{10}\,\mathrm{J\,cm}^{-3}$.

Table B.2 Energy densities of different plasmas
The short F & S refers to the publication on SL in sulphuric acid by Flannigan & Suslick [133]. MAGO (magnetic compression), JET (Joint European Torus), and NIF (National Ignition Facility, USA) refer to (fusion) plasma confinement machines and approaches described in more detail in chapter E.2. The numbers for the hydrogen bomb are based on Winterberg's temperature estimates [518]; in a radiation-dominated plasma the energy density is solely determined by the temperature.

quantity	T_i	T_e	n_i	n_e	u_{kin}	u_{rad}	u_{tot}
unit	[K]		[1/cm^3]			[J/cm^3]	
F & S	1.6e4		$n_e/3$	4e21	1.8e3	4.9e-5	1.8e3
MAGO	1.6e8	2.3e7	8e17		5.8e3	2.2e8	2.2e8
JET	1.7e8		3.3e19		2.4e5	6.9e11	6.9e11
NIF	5.2e7		1e25		4.3e10	5.6e9	4.9e10
sun	1.6e7		9.8e25		2.8e10	4.6e7	2.9e10
DT-bomb	2e8 - 8e8		–		–	2e12 - 3e14	–
DD-bomb	3.5e9		–		–	1.1e17	–

Table B.2 lists a few more examples of plasma energy densities. The calculation in the case of MAGO is based on the temperatures for ions and electrons of 10 and

2 keV and the number density given in [154]. For JET, the values come from Wesson [503] (p. 114), for NIF from [211].

So, from the smallest considered sound field to the most extreme listed plasma the energy densities cover an extremely wide and hardly imaginable range of 35 or perhaps almost 40 orders of magnitude. In order to anchor this scale we can still add two examples of chemical reactions: the fuel burn in a combustion engine and the explosion of TNT. Both are violent and forceful explosions, but we can still grasp their force based on everyday experience (or at least televised experience).

When burning octane in a car engine at a compression rate of 12:1 there are 3.3×10^{-3} g of oxygen molecules available corresponding to around 1×10^{-4} mol (and neglecting the volume fraction covered by the liquid fuel spray). Burning this at an ideal ratio together with 8×10^{-6} mol of octane releases an energy of 43.2 J inside one cubic centimetre of combustion flame.

How about TNT? The energy release by combustion is $\Delta H_c = (-3410 \pm 20) \times 10^{-3}$ kJ mol^{-1} [396]. With a density $\varrho = 1550\,\mathrm{kg\,m^{-3}}$ this implies an energy release per cubic centimetre of $23.3 \times 10^3\,\mathrm{J/cm^3}$. So we can see that the setup producing SL in sulphuric acid of Flannigan & Suslick approaches the same order of magnitude of energy density as TNT.

B.3 A note on twelve orders of magnitude

Barber et al. made a statement in the early nineties [26] that the degree of energy concentration can range up to twelve orders of magnitude. This was based on calculating the fraction of energy of the sound field shared by a single atom $\langle \varrho v^2 \rangle$ times volume per atom $\approx 4 \times 10^{-12}$ eV/atom and comparing it with the maximal energy $E = \hbar\omega = 6$ eV of photons they detected emanating from SL events of their experiment involving a helium bubble in water. This statement has since been cited from time to time (e. g. [11, 85, 276]) to underscore SL as an extraordinary research topic and an easy way to achieving hot and dense plasmas. However, also the belly of a firefly can provide molecules giving off photons of higher energy as orange glowing iron. And electric discharge sparks created with a polyester pullover is another interesting mechanism of energy concentration. The extraordinarity of SL hinges on the thermodynamics because light from chemical reactions or electrical discharge is less impressive than blue or UV photon emission from adiabatically compression-heated plasma. The 200 nm photons reported by Barber et al. with their energy of around 6 eV, if equated with the thermal energy kT would correspond to a temperature of over 66×10^3 K. However, maximal photon energy data generally carries less information than a whole spectrum recording does. Today, it is state of the art to interprete spectroscopic data from SL by fitting detailed particle interaction-based models [65, 187, 528, 529] and it is often stated that these models yield spectra with substantial deviations from a simple blackbody model.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
c	speed of sound/light
E	energy
F	force
H	enthalpy
$\hbar$	Planck constant (“quantum of action”)
K	bulk modulus
k	Boltzmann constant; spring constant
L_p	sound pressure level
n, n_e, n_i	particle/electron/ion density
p	pressure
T	temperature
T_e, T_i, T_r	temperature of the electron/ion/photon population
U	internal energy
u, u_{tot}	internal energy per volume, e. g. plasma energy density
u_e, u_i, u_r	internal energy per volume of electron/ion/photon population
u_{kin}	contribution of particle kinetic energy to the internal energy per volume
u_r, u_{rad}	contribution of photon energy to the internal energy per volume
V	volume
v	velocity
x	position, spatial coordinate
Z	degree of ionisation

List of Greek quantity symbols

Symbol	Description
σ	Stefan constant
ϱ	density
ω	frequency, e. g. of electromagnetic radiation

List of abbreviations

Abbreviation	Description
JET	Joint European Torus
MAGO	magnetic compression (магнитное обжатие, МАГО)
NIF	National Ignition Facility
SBSL	single-bubble sonoluminescence
SF	sonofusion
SL	sonoluminescence

Appendix C

The Rayleigh-Plesset equation

Due to its compressibility a gas bubble in a liquid can be excited into oscillatory radial motion. The gas may be considered to be the spring and the surrounding liquid the mass. The oscillation can be excited by an external harmonic sound field. The equation of motion ($F = ma$) of this system is the Rayleigh-Plesset (RP) equation:

$$R\dot{R} + \frac{3}{2}\dot{R}^2 = \frac{1}{\varrho}\left(p_l(t) - p_0 - p_a(t)\right) + O(\dot{R}/c), \tag{C.1}$$

where R is the bubble radius, p_0 the ambient pressure, p_l the pressure in the liquid just outside the bubble, p_a the acoustic pressure, ϱ the density of the liquid, and c its speed of sound. Several different ways[1] of deriving the equation are outlined and compared in [259], the most important equations are repeated here. Assuming a uniform gas pressure, the bubble's internal pressure p_i can be decomposed into vapour and (noncondensible) gas pressure and is offset with respect to the external pressure due to the surface tension σ, therefore

$$p_i = p_v + p_g = p_l + p_\sigma \qquad \Rightarrow \qquad p_l = p_v + p_g - p_\sigma. \tag{C.2}$$

The pressure rise in the compressed bubble can be expressed in terms of the equilibrium pressure $p_{g,e}$ as

$$p_g = p_{g,e}\left(\frac{R_0}{R}\right)^{3\kappa} = \left(p_0 + \frac{2\sigma}{R_0} - p_v\right)\left(\frac{R_0}{R}\right)^{3\kappa}, \tag{C.3}$$

whereby κ is the polytropic index which can be set to unity to represent the assumption of a bubble content in permanent thermal equilibrium with the surrounding liquid or to $\gamma = C_p/C_V$, the adiabatic index, for modelling completely adiabatic conditions. By accounting for damping through the liquid's viscosity η the interface pressure equilibrium becomes:

$$p_l = p_i - \frac{2\sigma}{R} - \frac{4\eta\dot{R}}{R} \tag{C.4}$$

[1] Another derivation is given in [390]. RP equations in different forms are also explained in [56, 257, 276].

which brings the RP equation into the most commonly seen form:

$$R\dot{R} + \frac{3}{2}\dot{R}^2 = \frac{1}{\varrho}\left(\left(p_0 + \frac{2\sigma}{R_0} - p_v\right)\left(\frac{R_0}{R}\right)^{3\kappa} + p_v - \frac{2\sigma}{R} - \frac{4\eta\dot{R}}{R} - p_0 - p_a(t)\right). \quad \text{(C.5)}$$

However, accounting only for viscous losses is often not realistic enough. Löfstedt et al. [276] managed to tightly reproduce a measured bubble size time series by using an extended form of the RP equation,

$$R\dot{R} + \frac{3}{2}\dot{R}^2 = \frac{1}{\varrho}\left(p_g(R) - p_0 + p_a(t) + \frac{R}{c}\frac{d}{dt}(p_g(R) + p_a(t))\right) \quad \text{(C.6)}$$

by including as the last term the energy loss of the bubble due to sound radiation into the surrounding liquid (the *Keller-Miksis* model [227][2]). Besides that, the adiabatic equation of state for the gas

$$p_g(R) = \frac{p_0 R_0^{3\gamma}}{(R^3 - a^3)^\gamma} \quad \text{(C.7)}$$

is modified such that a Van-der-Waals hard core radius a is accounted for, which will shift the minimal radius upwards.

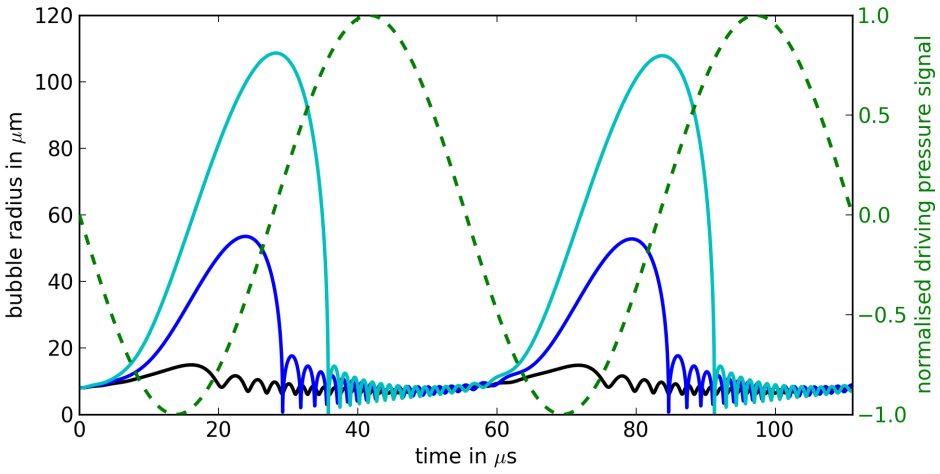


Figure C.1 Solutions of the Rayleigh-Plesset equation
The diagram shows solutions of the Rayleigh-Plesset equation for a bubble in acetone with an equilibrium radius R_0 of 8 µm driven by a sinusoidal external sound field at 18 kHz. The computations were made solely for illustrative purposes. The three traces in black, blue, and cyan correspond to different and increasing acoustic pressure levels of 0.85, 1.1, and 1.4 bar. The RP equation was simulated in the form of eq. C.5 with an additional simplified sound radiation term $(R/c)\dot{p}_g$. A fourth order Runge-Kutta scheme with R-dependent time-stepping was used for time-integration (the code can be found at [431]). The assumed properties of the liquid were $\sigma = 28\,\mathrm{mN\,m^{-1}}$, $\varrho = 790\,\mathrm{kg\,m^{-3}}$, $\eta = 0.4\,\mathrm{mPa\,s^{-1}}$, $c = 1174\,\mathrm{m/s}$, and for the bubble content $p_v = 0.093\,\mathrm{bar}$, $\kappa = 1.4$. Three important consequences of the increasing sound pressure level can be seen: it leads to (a) a larger bubble size as initial condition for the collapse, to (b) a delayed time of collapse under higher external pressure due to inertia, and consequentially to (c) higher final speeds (5.6, 2.1×10^3, and $22 \times 10^3\,\mathrm{m\,s^{-1}}$) of the bubble interface.

[2] [257, 276] contain brief discussions of the additional term

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
a	Van-der-Waals hard core radius
C_p, C_V	heat capacity at constant pressure/volume
c	speed of sound
p	pressure
R	bubble radius
t	time

List of Greek quantity symbols

Symbol	Description
η	dynamic viscosity
γ	adiabatic index
κ	polytropic index
ϱ	density
σ	surface tension

List of abbreviations

Abbreviation	Description
RP	Rayleigh-Plesset
SL	sonoluminescence

Appendix D

Detail descriptions for the SF experiment by Rusi P. Taleyarkhan et al.

The peer-reviewed journal publications of Taleyarkhan et al. concentrate on interpreting measurement results gathered during their SF experiments. The setup and preparation of their type of SF experiment is covered in less detail. Detailed descriptions of the used resonator apparatus including schematic drawings and the experimental protocol were disclosed by Taleyarkhan & West [464, 467] and Taleyarkhan [461, 462] in the form of patents. These descriptions and drawings seem to resemble the information given by Taleyarkhan to the RPI group for supporting them in their attempt to reproduce the ORNL SF experiment.

According to these descriptions, the following matrix of experiments is suggested (normal acetone is abbreviated as n-acetone, deuterated acetone as d-acetone):

- **d-acetone, external pulsed neutrons on & cavitation on:** This is the SF setup and when it works it should yield a statistically significant D-D fusion neutron (max. kinetic energy 2.45 MeV) and tritium measurement signature.

- **d-acetone, external pulsed neutrons on & cavitation off:** This control experiment should show that the irradiation by the pulsed neutron source (utilised for cavitation bubble nucleation) alone does not lead to the neutron and tritium detection signatures interpreted as evidence for SF.

- **n-acetone, external pulsed neutrons on & cavitation on:** This control experiment where everything is the same except that fusion fuel is missing should show that the SF detection signature is not correlated to a misinterpretable side-effect of the cavitation-inducing machinery being active.

- **n-acetone, external pulsed neutrons on & cavitation off:** This control experiment should show that there is no influence on the SF neutron and tritium detection outcome at all in connection with turning on the cavitation-inducing machinery.

The experimental **protocol for preparing the cavitation test section and the liquid** can be rewritten in the following form:

- The liquid (acetone) is filtered with a 0.5 µm filter "such as a coffee filter" [461].
- The test section is filled with liquid "such that the top reflector's piston dips into the fluid by about 5-6 mm under room temperature conditions" [461].
- For closing the test section the top head is sealed to the lower part of the glass resonator with a bead of RTV silicone [462].
- In [462] the top piston is described as free-floating on the liquid's surface. It is kept centre-aligned by employing a stainless steel guiding wire and capping its upper tube or opening with a "thin acrylic or other such disc" (presumably with a hole in it so as to allow vertical sliding along the guiding wire). However, figures 2a & 2b of [464] may show the same or an alternative, perhaps non-sliding, wire-based support of the upper piston.
- Tightening of the vacuum system; evacuation, lowering the pressure close to the vapour pressure of the liquid which is cooled to 0-5 °C.
- Degassing with the help of acoustic agitation (the frequency may be different from the resonance frequency): the liquid bubbles up or even becomes "foamy". Further pumping is needed to maintain the bubbly state and support the degassing process.
- When the bubbling due to degassing dies out the agitated resonator steadily turns into a neutron detector, further bubbling can only be triggered by neutron scattering if sufficiently large states of tension are reached in the liquid. Therefore it is necessary that the agitation now matches the resonance frequency. The driving amplitude needs to be adjusted upwards following the process of further and further degassing. Neutron scattering-triggered bubbling comes in the form of "bubble jets" (also referred to as "comets" or "streamers") originating at the sound pressure antinode and pointing outwards. The lifetime of the jets decreases. When no more jets appear but only bursts of bubble clusters then the liquid is sufficiently degassed.
- In a variation of the experiment dissolved alpha emitters (e. g. uranium salt) can be used for bubble nucleation rather than using an external neutron source.
- In the case of using an external neutron source the working mode as neutron detector can be readily checked by switching the neutrons source on and off (or: taking it away and bringing it back) and observing that no more cavitation occurs without the neutron source.
- When the "comet-like structures turn into spherically looking bubble clusters" [461] then the optimal working regime of SF experiments has been reached.

- In acetone a sound pressure calibration procedure is proposed. It is known that the sound pressure threshold for neutron-induced cavitation in cooled acetone lies around -7 bar. By slowly increasing or reducing again the amplitude of the voltage signal supplying the transducer the cavitation threshold can be found. Doubling of the voltage amplitude was the procedure employed by Taleyarkhan et al. for achieving the intended regime of circa -15 bar (or up to -20 bar [464]) peak tension in the cooled acetone. The procedure relies on the assumption that the electronics, the transducer, and the resonator are not too far away from operating in a linear regime.
- In the case of using an external pulsed neutron source (PNG), the steps for determining the driving voltage corresponding to the onset of neutron-induced cavitation can be repeated systematically for different phase offsets between acoustic drive and neutron pulses. It can serve as a straightforward procedure for synchronising the neutron pulses with times of maximum tension (see paragraph [0080] in [464]). It could be added that as an additional advantage this approach avoids the need to gain precise knowledge about the phase angle offset of any sound pressure measurement device such as a hydrophone.
- According to [464] (paragraph [0091]), the height of the acetone filling level "should be carefully selected". One can infer that this setting had been identified as one of the more sensitive parameters.

Some knowledge of and familiarity with the experimental procedures at RPI have surely influenced and hopefully improved my formulation of the above experimental details.

A decisive additional piece of information supplied from Taleyarkhan to the RPI team on how to conduct the SF experiment is the resonator sketch shown in figure D.1. It closely resembles figures 2a & 2b of [464].

List of abbreviations

Abbreviation	**Description**
ORNL	Oak Ridge National Laboratory
PNG	pulsed neutron generator
RPI	Rensselaer Polytechnic Institute
RTV	room temperature-vulcanising (silicone)
SF	sonofusion
SL	sonoluminescence

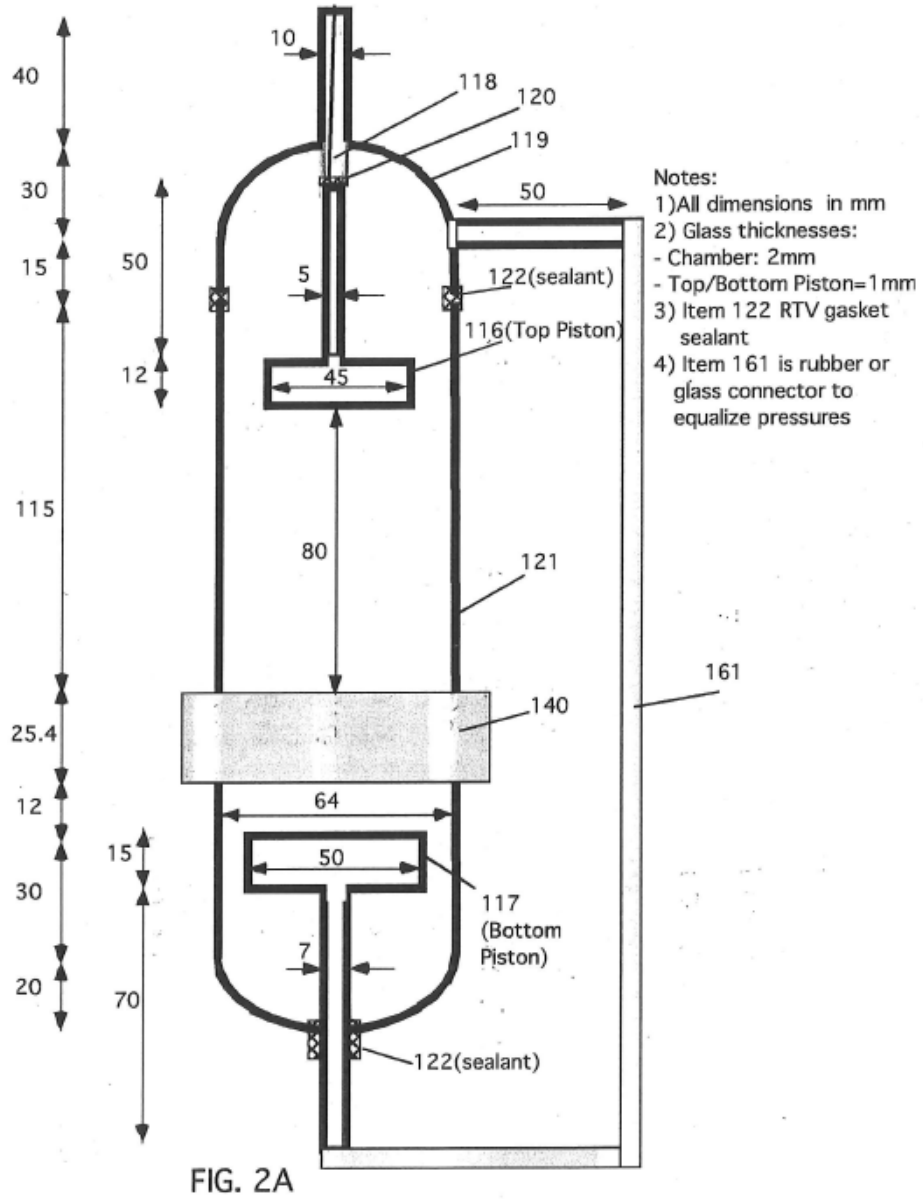


Figure D.1 Taleyarkhan's resonator sketch.
Some notable details that can be inferred from the sketch: 1) The hollow glass pistons have different diameters, the bottom one is wider. 2) The pistons' interior volumes are pressure-equilibrated with the rest of the chamber. 3) The RTV bead sealing the top head leaves a distance between the glass parts on the order of the glass wall thickness. 4) The gap bridged by the RTV bead sealing the vessel's bottom outlet and fixating the bottom piston is either intended to be very narrow or it had not been deemed important to give any hint on the width of the gap between the glass parts. 5) A rather square and edgy cross section of the pistons had been deemed ideal or the exact shape of the pistons had been assumed to be of minor importance. (Reprint with permission of R. Taleyarkhan.)

Appendix E

Some basic physics of nuclear fusion and plasma confinement

Going out from what nuclear fusion is and what is necessary to make it happen, different mechanisms are explained by which nature and humans can trigger fusion reactions. These mechanisms are basically mechanisms of energy concentration. Putting the sonofusion (SF) experiment in a general context means comparing the energy concentration mechanism of sonoluminescence (SL) with the other ones.

E.1 Nuclear fusion: the basics

E.1.1 Nuclear binding energies

Chemical reactions happen when molecules are modified, when atoms are regrouped into different molecules than before. When the resulting molecules contain a lower amount of potential energy than the initial molecules, then the reaction is exothermic, otherwise endothermic. The burning of coal is a good example of an exothermic reaction, where the energy investment needed for breaking bonds among C atoms and O_2 molecules is less than the energy gain from forming CO_2 from free atoms. One can also speak of *binding energies* and say that the sum over the binding energies of the two C−O bonds in carbon dioxide is larger than the binding energies involving one oxygen molecule and one carbon atom in the middle of a coal block. By definition, the binding energy is the energy gained when taking two particles being far apart and bringing them close together so they are in their stable equilibrium position, i. e. each particle is at the bottom inside the potential well created by its neighbour's presence. Knowing about the binding energies of molecules allows to determine whether chemical reactions are endothermic or exothermic. Secondly, thinking in terms of binding energies often helps to understand how large the energy barrier is which has to be overcome for a chemical reaction to happen, i. e. how far coal and oxygen need to be heated up so the molecules break.

Analogously, knowing about the binding energies of atomic nuclei helps to understand nuclear reactions. Nuclear binding energies are determined by the laws of quantum physics, and in that framework they can be computed with a high precision. Yet, a simple and in many cases efficient formula for calculating the binding

energy of a nucleus with a total number of A nucleons of which Z are protons is the *Bethe-Weizsäcker formula*, also called the *semi-empirical mass formula*:

$$E_B = a_V A - a_S A^{\frac{2}{3}} - a_C \frac{Z^2}{A^{\frac{1}{3}}} - a_A \frac{(A-2Z)^2}{A} - \delta(A,Z) \tag{E.1}$$

with

$$\delta(A,Z) = \begin{cases} -\delta_0 \cdot A^{-\frac{1}{2}} & \text{if } Z \text{ and } N \text{ both even,} \\ 0 & \text{if } A \text{ odd,} \\ +\delta_0 \cdot A^{-\frac{1}{2}} & \text{if } Z \text{ and } N \text{ both odd.} \end{cases}$$

The five contributions to this sum represent volume, surface, Coulomb, asymmetry, and pairing energies. The first two terms are the ingredients of the droplet model, the next one is the electrostatic repulsion of the protons, and the last two terms represent rules arising from the rules for filling up the quantum-mechanical energy levels available in the core's potential well for the two sorts of nucleons.[1] Figure E.1 shows the binding energies per nucleon computed with this formula overlaid with empirical data for stable isotopes and uranium. It can be seen that the elements around iron and nickel exhibit the highest binding energies, in particular $^{62}_{28}$Ni has the highest binding energy per nucleon, and that exothermic regroupings of nucleons have to go from left to right if going out from light nuclei, and that they have to go from right to left when the reactants are heavier than nickel. The latter process of splitting heavy isotopes into lighter ones is called *nuclear fission*, while the process of merging light nuclei into heavier ones is called *nuclear fusion*. Table E.1 lists some fusion reactions of relevance in the context of humans triggering fusion in various earthly laboratory environments. (These achievable fusion reactions can be contrasted with an extremely improbable one, the one of breeding neutrons going out from only protons (listed at the bottom of table E.1). If the cross section of this reaction was not so extremely small[2], then it would not be possible for stars to keep burning over billions of years [88].)

E.1.2 Why are fusion reactors more difficult to build than fission reactors?

A uranium atom in a fission reactor does not need to be fast, even the neutrons carrying the fission chain reaction along do not need to be fast. When a ^{235}U core captures a neutron and undergoes fission, then the nucleus is split into two droplets of the core matter, the two daughter nuclei. At that moment the strong nuclear force between them has become too weak to be dominant any longer. Now the electrostatic repulsion between the two nuclei with their positive charges is at play and accelerates the two cores away from each other. The fast cores transfer their kinetic energy through collisions to other atoms until the nuclei are thermalised, i. e. the kinetic energies have been distributed evenly. As neutrons are insensitive to

[1] Appendix F gives more detailed descriptions of the physical background of these terms and expands the view that nuclear binding energy gradients are the main drivers of our universe's energy houshold.

[2] According to [88], a thick hydrogen target would have to be bombarded for ten years with 1 Ampère of protons at 1 MeV to obtain just one reaction.

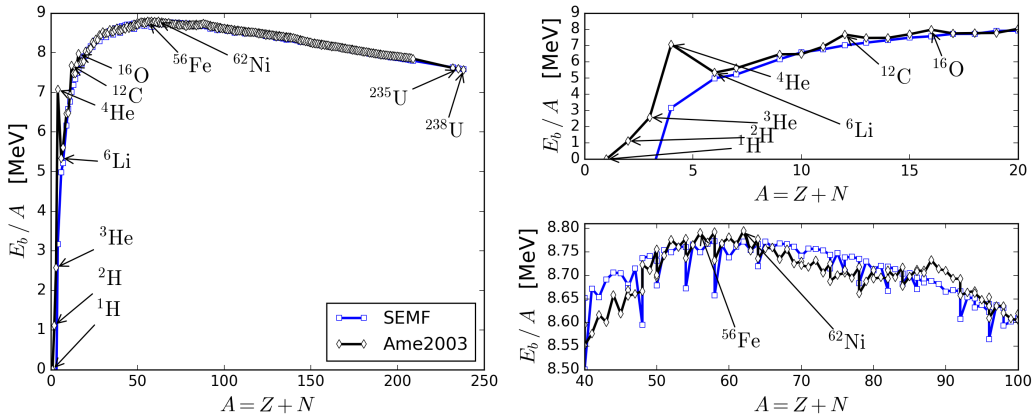


Figure E.1 Nuclear binding energies of stable isotopes.
Atomic core configurations are characterised by the number of protons Z, the number of neutrons N, and the total number of nucleons $A = Z + N$. The binding energy E_b is the energy gained when forming a compound from free (far apart) particles. The diagrams show the binding energy divided by the number of nucleons. Energy values gained by evaluating the Bethe-Weizsäcker formula are depicted in blue. The coefficients for the semi-empirical mass formula (SEMF) in MeV were taken from Chowdhury & Basu [86]: $a_V = 15.409$, $a_S = 16.873$, $a_C = 0.695$, $a_S = 22.435$, $\delta_0 = 11.155$. These binding energy values are compared to experimentally gained numbers compiled by Audi et al. [16], labelled "Ame2003". The stable isotope listing was adopted from Skyman [421, 422]. Among stable isotopes $^{62}_{28}$Ni is the one with the highest binding energy per nucleon [131]. That means energy can be gained by regrouping nuclear matter into heavier units going out from nuclei lighter than nickel, or by regrouping into lighter units starting from heavier isotopes. The first process is called nuclear fusion, the latter one fission. Although having no stable isotopes, uranium is added into the diagram because $^{235}_{92}$U is the primary fuel for civil-use nuclear fission power plants.

Table E.1 Some important fusion reactions

Reactions relevant for laboratory fusion		
D + T	→	^{4}He (3.52 MeV) + n (14.06 MeV)
D + D	→	T (1.01 MeV) + p (3.03 MeV)
	→	^{3}He (0.82 MeV) + n (2.45 MeV)
D + ^{3}He	→	^{4}He (3.67 MeV) + p (14.76 MeV)
T + T	→	^{4}He + n + n (11.32 MeV)
^{3}He + T	→	^{4}He + p + n (12.1 MeV)
	→	^{4}He (4.8 MeV) + D (9.5 MeV)
	→	^{5}He (2.4 MeV) + p (11.9 MeV)
p + ^{6}Li	→	^{4}He (1.7 MeV) + ^{3}He (2.3 MeV)
p + ^{7}Li	→	2^{4}He (22.4 MeV)
D + ^{6}Li	→	2^{4}He (22.4 MeV)
p + ^{11}B	→	3^{4}He (8.682 MeV)
n + ^{6}Li	→	^{4}He (2.1 MeV) + T (2.7 MeV)
The bottleneck reaction of fusion burn in stars [88]		
$p + p$	→	D + $e^+ + \nu$

the electrostatic field, and since they can approach an atomic nucleus at low speed, a fission chain reaction can be sustained also while keeping the reactor at room temperature, which is the case for some types of research reactors. The thing is that

a bulk of the fission reactor can be kept at a thermal equilibrium of much lower temperature than the temperature represented by the small fraction of particles which are fast and carry energy gained from nuclear reactions.

If one wants to induce fusion reactions the situation looks different: the reactants are both positive, repel each other, and need to be brought to collision at high enough speed, so they approach each other closely enough and the strong nuclear force kicks in. In principle there are two ways of supplying the collision speed, the use of an accelerator or the use of high temperatures.

Why can a fusion reactor for electricity production not be based on the accelerator principle? This has to do with thermodynamics, entropy, and what it means for energy input, confinement, and losses in a setup with directional particle movement (accelerator) versus random particle motion (thermal equilibrium). The argument is laid out very nicely by A. Piel in [351]. With an accelerator one can produce a particle beam of e. g. deuterium or tritium ions and focus it on a target which we would for the moment assume to be a cube of solid deuterium ice. Each incoming ion has two ways to interact with the target matter: with every nucleus and each electron it can interact over the electromagnetic field or with the nuclei via the short-range strong nuclear force (effective over only $\sim 1 \times 10^{-15}\,\mathrm{m}$ [347]). Only the latter interaction can lead to the nuclei catching each other and fusing. The problem is that interaction with the $\vec{E}$-field, the Coulomb interaction, is possible throughout the ice cube, but direct core-core collisions are only possible upon direct hits, and assuming a tritium ion flying straight towards a centimetre-size ice-cube, only behind 0.027 % of its front surface lie those impact paths that lead to a direct knock-on collision among nuclei. Will it help to put many ice cubes behind each other into the beamline? No, because the Coulomb scattering will stop the beam and turn most of its energy into heat, long before the tritons have flown through enough deuterium ice to have seen much higher decent direct hit probabilities. Fusion does produce a lot of energy, but by far not enough to be able to afford the necessary loss of a realistically large fraction of accelerated ions to Coulomb interaction.[3] The resulting dilemma is: if the target material is being kept stable by cooling, then most of the energy investment for accelerating the ion beam is lost in target heating, but if you let the target heat up, so the energy stays largely inside it and can ultimately contribute to fusion reactions, then you have transitioned to the second approach, the thermonuclear approach.

The other approach, thermonuclear fusion, consists in provoking fusion reactions triggered by unordered random collisions in a plasma of sufficiently high temperature. And with that approach we are back to the coal flame which cannot be cold because in order to sustain the flame the reaction products have to transfer their energy to the reactants through a thermodynamic equilibrium. From here arises the question of what type of vessel should be used to contain the extremely hot fusion plasma? What construction can bear the necessary pressures and temperatures and

[3] The device examined in [294] poses an interesting challenge to the above critique of accelerator-like reactor designs based on fundamental thermodynamical considerations. It is discussed below on page 244 in the paragraph on inertial-electrostatic confinement. In that device the Coulomb interaction ("the very forces that usually thermalise [a non-Maxwellian non-neutral plasma]" [294]) stabilises oscillating ion bunches.

how long are the fusion conditions to be maintained?

In combustion devices like coal power plants, turbines, or piston engines, the structural materials are always kept at substantially colder temperatures than the flame. The burning gas is pushing mechanical parts like blades and pistons, and there are still layers of cooler gasses in between which transfer the work and the pressure. So where lies the problem with the fusion flame? The problem is on the one hand one of orders of magnitude and on the other hand one of different mechanisms of energy transfer. First about the orders of magnitude concerning temperature and pressure. The temperature is correlated directly to the particle kinetic energies via $E = kT$. The temperatures allowing substantial rates of D-D or D-T fusion are beyond 10^7 K. The pressures, however, can range from very low, e. g. 2×10^{-5} mbar for the JET Tokamak reactor [128], to quite high, e. g. 2.5×10^{11} bar at the centre of the sun. The other thing is that heat transfer by radiation plays a minor role at the temperatures of combustion flames, and the heat transfer via conduction and convection is slow enough from the flame to the metal parts (in a piston engine or turbine), so it does not take away too much of the heat of the expanding and mechanically working gas during the relevant time scale. But this is different for small-sized plasmas at 10^8 K where radiation leads to substantial heat losses on much shorter time scales.

Therefore, the question of the techniques used for *plasma confinement* is a central one in the quest for a machine capable of producing controlled thermonuclear fusion, i. e. how can the matter and the energy which form the plasma be brought and kept together, and how long can fusion conditions be kept up? It makes sense to classify fusion plasma experiments according to their confinement methods. The fundamental choices are: *gravitation confinement*, *inertial confinement*, and *magnetic confinement*. In a short list of exemplary fusion devices presented below (section E.2), it is described how the confinement mechanisms play out in practice.

E.1.3 Plasma basics

Plasma: equation of state

A plasma is a gas with three components: ions, electrons, and photons. For many thermodynamical aspects the assumptions of an ideal gas can be made. In an ideal gas of one particle type the equation of state can be written as

$$pV = NkT \qquad \text{or} \qquad p = nkT \tag{E.2}$$

with N the absolute particle number in the volume of interest, V the volume, and the particle number density $n = N/V$. Next to this thermal equation of state, the caloric equation of state has the form

$$U = \frac{3}{2}NkT \qquad \text{or} \qquad u = \frac{U}{V} = \frac{3}{2}nkT \tag{E.3}$$

which is founded in the relation

$$U = N\langle E_{\text{kin}}\rangle \qquad \text{with} \quad \langle E_{\text{kin}}\rangle = \frac{m}{2}\langle \vec{v}^2\rangle = \frac{3}{2}kT. \tag{E.4}$$

In a plasma the corresponding equation of state looks a bit more complex, because of the multiple species, but also, because in general there is a contribution to the pressure arising from the momentum carried by photons. This pressure contribution p_r is called *radiation pressure* and it is similarly connected with the energy density u_r of a photon gas (in thermal equilibrium), which is a function of the temperature only, by these two equations:

$$u_r = \frac{4\sigma}{c}T^4 = aT^4 \qquad \text{and} \qquad p_r = \frac{4\sigma}{3c}T^4 = \frac{a}{3}T^4 = \frac{u_r}{3}. \tag{E.5}$$

Here, σ is the Stefan constant

$$\sigma = \frac{2\pi^5 k^4}{15c^2h^3} = 5.670\,400 \times 10^{-8}\,\mathrm{J/(sm^2K^4)} \tag{E.6}$$

and the factor $a = 4\sigma/c$ is called the radiation constant

$$a = \frac{4\sigma}{c} = 7.565\,767 \times 10^{-16}\,\mathrm{J/(m^3K^4)}. \tag{E.7}$$

For the plasma with its multiple particle species the energy density is the sum of the energy densities of the three subsystems

$$u = \frac{U}{V} = u_i + u_e + u_r = \frac{3}{2}n_i kT_i + \frac{3}{2}n_e kT_e + \frac{4\sigma}{c}T_r^4 \tag{E.8}$$

and accordingly the pressure is

$$p = p_i + p_e + p_r = n_i kT_i + n_e kT_e + \frac{4\sigma}{3c}T_r^4. \tag{E.9}$$

In a Z times ionised[4] gas in temperature equilibrium and if radiation is negligible, the expressions can become very simple

$$p = (1+Z)nkT, \qquad u = \frac{3}{2}(1+Z)nkT. \tag{E.10}$$

And some more useful equations can be gained by writing the energy density as a product of concentration n, atomic weight A, the proton mass m_p, and with the help of the plasma density $\varrho = nAm_p$ and the specific heat capacity at constant volume c_V

$$u = \varrho c_V T \qquad \text{with} \qquad c_V = \frac{3(1+Z)k}{2Am_p}. \tag{E.11}$$

A simplified writing of the equation of state can then be

$$\frac{p}{\varrho} = \frac{2}{3}c_V T. \tag{E.12}$$

[4]In a partially ionised gas the probabilities of various degrees of ionisation can be calculated with the Saha equation. The Saha equation is based on comparing ionisation and recombination rates with Boltzmann factors. The key detail is that also induced recombination, a three-body-process, is taken into account. For fully ionised gasses it is not of relevance.

The above-discussed case of negligible radiation contributions can be compared to the other extreme of high temperatures, where the radiation terms with T^4 dominate all other contributions and which leads to

$$u = aT^4 \qquad \text{and} \qquad p = \frac{u}{3}. \tag{E.13}$$

Then pT^{-4} is constant. The kinetic gas theory says that for an ideal gas during reversible adiabatic changes the term $pT^{-\gamma/(\gamma-1)}$ stays constant, where $\gamma = c_P/c_V$ is the specific heat ratio. So, for e. g. a mono-atomic gas, γ transitions from $\gamma = 5/3$ at low temperatures to $\gamma = 4/3$ in the radiation-dominated regime [518]. The mean free path of a photon can also be taken as a criterion distinguishing the two cases. If the dimensions of the plasma are large compared to it then it is radiation-dominated [518].

Plasma: shielding and other force competitions

In the plasma the negative and positive charge carriers are moving freely, and these movements would in principle equilibrate any electric potential gradient. But the electric field as driver of particle motion is in competition with gas kinetics and its tendency to equilibrate concentration gradients. As a result the characteristic length scale, outside which two charge carriers are completely shielded by the surrounding plasma and cannot interact anymore, the Debye length λ_D

$$\lambda_D = \sqrt{\frac{\epsilon_0 k T_e}{q^2 n_e}} \tag{E.14}$$

grows with the temperature.[5]

A key value for the competition of thermodynamic and magnetic forces is the value β which is defined as a pressure ratio:

$$\beta = \frac{nkT}{B^2/(2\mu_0)} \tag{E.15}$$

is the ratio between thermodynamic and magnetic pressure. At low β (i. e. $\beta \ll 1$) particles are sparse, they move under the influence of a given background pattern of magnetic fields. For denser plasmas a skin of currents can completely decouple[6] the magnetic fields inside and outside a plasma core region. If the skin currents are established in such a way as to push all magnetic field lines out of the core region creating a field-free volume where charged particles fly in straight paths in between collisions, and if the driving force leading to the establishment of that current pattern is the thermodynamic tendency to maximise entropy, to destroy order, the tendency of the particle ensemble to get to an isotropic distribution of motion directions, then one speaks of a plasma with high β, with $\beta \approx 1$. The scenario can be imagined with

[5] In [351] another helpful picture is used for imagining temperature as a force hindering the shielding mechanism. Higher kinetic energies make the trajectories "stiffer" (so they show less curvature at a given field strength) so the orbits cannot wrap around opposing charges as narrowly.

[6] examples explained below: polywell, FRCs, spheromaks

the plasma being capable of withstanding the tendency of the external magnetic field to spread into the plasma.

Another consequence of the electron and the ion systems consisting both of freely moving particle types which attract each other is that they can oscillate against each other. The frequency of their experimentally observable natural collective mode is close to the expression below which is usually called the plasma frequency[7],

$$\omega_p = \sqrt{\frac{n_e e^2}{m_e \epsilon_0}} = \frac{v_e}{\lambda_D} \tag{E.16}$$

and which is the reciprocal of a time scale $\tau \approx \lambda_D / v_e$ with $v_e \approx \sqrt{kT_e/m_e}$. The meaning of that τ can be understood as the time scale needed for the electron system to reach a new equilibrium distribution after perturbations of the potential landscape under the assumption that the amplitude of the potential perturbation is small compared to the average kinetic energy of the electrons and thus speed variations are also small.

The reason why usually in the definitions of the Debye length and the plasma frequency only the properties of electrons show up, and not the characteristics of the ion system, is the mass difference. Since $m_p/m_e \approx 1836$, the electrons have by a factor ~ 43 the higher average speeds and react much quicker to accelerating forces as compared to the lightest ions. Therefore, many characteristic plasma quantities are deduced under the assumption that the ions are fixed in space while the electrons equilibrate electric potential gradients.[8]

Plasma: limits of the ideal gas and collectivity assumptions

There are several important criteria telling whether a plasma is beyond the validity range of the assumptions of the kinetic gas theory. The first one is the *coupling parameter* [347, 351]

$$\Gamma = \frac{E_{\text{pot}}}{kT} = \frac{Z^2 e^2}{akT} \tag{E.17}$$

The particle spacing parameter a used in the expression above is the radius of a sphere

$$a = \sqrt[3]{\frac{3}{4\pi n_i}}. \tag{E.18}$$

where the sphere contains the volume available per ion. The radius a is also called the Wigner-Seitz radius. For $\Gamma \gtrsim 1$ the potential energy dominates over the kinetic energy, and one speaks of a *strongly coupled plasma*.

Another criterion based on spacing is the *plasma parameter* N_D [146, 351], defined e. g. for electrons by

$$N_{De} = \frac{4\pi}{3} \lambda_{De}^3 n_e. \tag{E.19}$$

[7] To be more precise one could insist on the term *electron plasma frequency* and label it $\omega_{pe} = v_e/\lambda_{De}$.

[8] A more exact and general treatment is e. g. the definition of two Debye lengths λ_{Di} and λ_{De} for ions and electrons with the corresponding masses and temperatures and computing λ_D through $1/\lambda_D = 1/\lambda_{Di} + 1/\lambda_{De}$ [351].

The plasma parameter N_{De} must be $\gg 1$ for a plasma to be weakly coupled. This ensures that there are many electrons within a Debye sphere and the consequence can be expressed as long-range collective effects dominating short-range Coulomb interactions [146], or equivalently, that Debye shielding is a collective process and that a statistical justification of its length scale is justified [351].

The final criterion to be mentioned here distinguished whether quantum effects dominate the behaviour of the gas or not. Quantum effects begin to play a role once the inter-particle distances are reduced to a size comparable with their thermal de Broglie wavelengths

$$\lambda_B = \frac{h}{m_e v_{Te}} \qquad \text{with} \qquad v_{Te} = \sqrt{\frac{2kT_e}{m_e}}. \tag{E.20}$$

This happens first for the electrons because their wave functions are spread out much farther. Once their wave functions overlap, electrons as fermions are forced to occupy different quantum mechanical states due to the Pauli exclusion principle. This ensues the transition from Boltzmann to Fermi-Dirac statistics. The electron plasma in metals has to be treated that way.

Plasma waves

Diverse sorts of waves can travel through a plasma, e. g. acoustic waves of the electron or the ion subsystem, or electromagnetic waves. These waves are subject to varying and characteristic degrees of damping, mainly because of the interactions of the subsystems through collisions. One can think of friction between the subsystems in unequal motion. The inequality of the acoustic wave motions as a consequence of differing electron and ion masses means that acoustic waves induce longitudinal electrostatic waves. A second damping mechanism, *Landau damping*, is due to electrons being accelerated by longitudinal electrostatic waves [303, 351]. The damping mechanisms are the basis for plasma heating through electromagnetic waves.

The multiple-species nature of the plasma and the sliding out of phase of acoustic waves has strong consequences in particular for shock waves. The smaller particles with the faster thermal speeds are expelled from the shock wave and surf in front of it. This means shock fronts in plasma have a thermal precurser due to electrons surfing[9] ahead of the main ion density jump. In laser confinement fusion research this preheating effect has undesired consequences and its mitigation has become a particular interest (see section E.2.3).

The Lawson criterion

Thermonuclear fusion can only be achieved in a hot plasma. The production of the plasma represents a large energy investment that has to be payed upfront. The Lawson criterion has the purpose to determine whether this investment pays off [347, 351, 518]. It compares the energy cost of the plasma buildup with the fusion reaction enthalpy released in the plasma. Thus, for a given plasma in which the

[9] This is the same effect as seen in molecular dynamics simulations of sonoluminescing bubble implosions. See appendix A.1.6 and [30].

thresholds for fusion reactions are reached, it indicates the minimal confinement time after which the fusion energy surpasses the compression and heating cost. In the other direction, for a confinement technique with a characteristic confinement time constant τ the minimal fusion rate can be estimated beyond which net energy is gained. In the sense that the plasma decay time τ is taken as a measure of the losses (through conduction, radiation, expansion) the Lawson criterion can be seen as a comparison of energy gains and losses.

If the fusion reaction rate is $\mathcal{R}$ and the reaction enthalpy is Q, then the total energy output is

$$E_{\text{out}} = \mathcal{R}\tau Q. \tag{E.21}$$

In a pure DT plasma the concentrations of the two reactants is $n/2$ where n is the ion density in this case. The reaction rates can thus be written as

$$\mathcal{R}_{DT} = \frac{n^2}{4}\langle v\sigma_{\text{DT}}\rangle \qquad \text{and} \qquad \mathcal{R}_{DD} = n^2\langle v\sigma_{\text{DD}}\rangle \tag{E.22}$$

when expressing them in terms of a reaction cross section σ. Equating the energy that was necessary for heating and compression with the total kinetic energy

$$E_{\text{kin}} = \frac{3}{2}k(n_eT_e + n_iT_i) = 3nkT \tag{E.23}$$

the payoff condition means

$$\begin{aligned} E_{\text{kin}} &< E_{\text{out}} \\ 3nkT &< \frac{n^2}{4}\langle v\sigma_{\text{DT}}\rangle\tau Q_{\text{DT}} \quad \text{for DT,} \\ 3nkT &< n^2\langle v\sigma_{\text{DD}}\rangle\tau Q_{\text{DD}} \quad \text{for DD,} \end{aligned}$$

and this yields

$$n\tau > \frac{12kT}{\langle v\sigma_{\text{DT}}\rangle Q_{\text{DT}}} \quad \text{for DT,} \tag{E.24}$$

$$n\tau > \frac{3kT}{\langle v\sigma_{\text{DD}}\rangle Q_{\text{DD}}} \quad \text{for DD.} \tag{E.25}$$

The enthalpy Q and the temperature-dependent average nuclear reaction cross section ion velocity product $\langle v\sigma\rangle$ can be looked up from tabulated empirical data. Assuming a sphere of radius r for the relevant part of the plasma, and relating it to the disassembly time τ via

$$\tau = \frac{r}{a} \qquad \text{with the escape velocity} \quad a = \sqrt{\frac{3kT}{m}},$$

then the criterion can be transformed to evaluate the product nr instead of $n\tau$. And by replacing n with $N_{Av}\varrho/A$ in a next step, where N_{Av} is the Avogadro number and A the atomic mass number, it can be further changed to evaluate the product ϱr [518].

Sometimes, better suitable formulae for certain application cases can be gained by letting two more estimation factors flow in [518], an energy multiplication factor

$F = E_{\text{out}}/E_{\text{in}}$ and an energy conversion factor $\varepsilon = E_{\text{kin}}/E_{\text{in}}$. The latter factor gives the fraction of an initial energy input which really ends up as the thermal energy contained in the plasma volume of interest. Then, e. g. equation E.24 turns into

$$n\tau = \frac{F}{\varepsilon}\frac{12kT}{\langle v\sigma\rangle Q}.$$

It must be noted that the Lawson criterion defined as above can tell whether a burn with a net energy payoff has been created, and that the self-heating by the fusion products has been ignored so far. The next interesting question is whether the effect of self-heating can become strong enough to sustain a fusion detonation wave going out from the hot spot and burning through a surrounding fuel supply. On the small scales of laser confinement fusion and sonofusion the reaction enthalpy carried away by neutrons has to be left out and only ions can be taken into account, because the mean free path of the neutrons is much too large. A simple modification of the Lawson criterion along these lines can be found in [518].

E.2 Nuclear fusion: plasma confinement mechanisms

The central problem of energy harvesting through thermonuclear fusion is the question of how to contain the extremely hot fusion plasma in a controlled volume, and how to transfer its energy output to the peripheric machinery of a conventional (e. g. steam cycle) power plant. The machinery has to be prevented from being destroyed by heat and radiation, yet there has to be some level of contact for carrying gained heat away from the plasma. The question of fusion plasma confinement. The crucial question of plasma confinement is the question of the right amounts of matter and energy together over the right amount of time. The energy investment of creating and heating the plasma has to pay off; and with different techniques of plasma creation and different types of plasmas being produced the time scale for energy amortisation is of different length. The following pages are devoted to giving a broad overview and classification of solution approaches to solve this problem, i. e. of *plasma confinement methods*.

E.2.1 The sun (gravitational confinement)

The matter forming the sun is being kept together by the sun's gravitation field. In analogy to other confinement principles one can speak of *gravitational confinement*. The energy of the plasma inside the sun is kept together because the huge size of the sun has the consequence that energy produced at the centre of the sun needs a very long time to reach the surface. The energy leak rate, the amount of energy lost through radiation over a given time interval in relation to the total energy content, shrinks with a growing size of the plasma. Another question is: what heated the sun up initially and ignited the fusion fire? This came about by the collapse of the hydrogen cloud under its own gravitation which transformed the potential energy of the initial distribution of mass into heat.

The conditions at the core of the sun are: $T = 1.571 \times 10^7\,\text{K}$, $p = 2.477 \times 10^{11}\,\text{bar}$, and $\varrho = 1.622 \times 10^2\,\text{g cm}^{-3}$ [179]. The fusion process results in a heat production

of $276.5\,\mathrm{W\,m^{-3}}$ at the solar core [88][10], which is similar to the heat per volume generated in a compost heap [89, 243]. The time scales for heat transport from the core to the surface are the Kelvin-Helmholtz time scale[11] with 3×10^7 years and the the photon diffusion time scale[12] with 1.7×10^5 years [427]. Luckily, the $p + p$ fusion reaction has such an incredibly small cross section that it limits a star's heat generation to the level of compost, giving it such long-term stability and eventual biological evolution in its surroundings a chance to develop complexity.

E.2.2 The H-bomb (inertial confinement)

The plasma ball produced by the explosion of a hydrogen bomb in the atmosphere, one might say, is not confined at all, it is free to spread out. How can inertia be a confinement mechanism? In the case of the H-bomb the fusion plasma ball is surrounded by layers of high-density metals (large atomic number Z), and the fusion ignition process involves a compression movement. The pressure buildup in the fusion fuel has to reverse that inward movement and accelerate the heavy materials away from the centre. The higher the outside mass, the slower the acceleration. These devices work because they can be setup in such a way that the fusion burn time period, in which a substantial fraction of the fusion fuel is consumed, is comparable to or shorter than the time needed for reversing the compression movement of the heavy coating layer (*tamper*) and expanding it substantially [181, 517]. This so-called disassembly time[13] can be estimated to be around 5-20 ns. The plots in [181] show that the time from fusion ignition to 75 % burn-up can take between 2 and 20 ns.

Figure E.2 contains a sketch of the Teller-Ulam-Sakharov-Zel'dovich design of the H-bomb. The principle is that a fission bomb compresses and by that compression heats the fusion fuel until the fusion flame ignites. The black lines around the fission plasma (shaded blue) and around the fusion fuel (green) symbolise coatings of high-Z (tamper) material necessary for slowing down the reversal of the compression movement into expansion and the buildup of expansion speed.

What are the plasma conditions at fusion ignition and during the burn according to [181, 517]? The ignition temperatures are given as $5 \times 10^7\,\mathrm{K}$ (4 keV) for D-T and $5 \times 10^8\,\mathrm{K}$ (40 keV) for D-D plasma. These are hot spot ignition temperatures. In

[10] The value can be found in table 6-6 on page 483 of [88], which is a reprint of: Bengt Strömgren, *Stellar Models for Main-sequence Stars and Subdwarfs* in L. H. Aller and D. B. McLaughlin (eds.), *Stellar Structure*.

[11] The definition of this time scale is not built on any assumptions on a star's internals (also fusion power is ignored), but only on the observed energy loss, the luminosity. When the gas sphere contracts under gravitation, it heats up, and the energy loss responsible for cooling is known because a star's brightness and spectrum are the most readily available information. The Kelvin-Helmholtz time scale then determines how long it takes for the collapsing gas sphere to shrink and cool down and settle completely, and it is given by $\tau_{\mathrm{KH}} = U/L$, the potential energy divided by the luminosity.

[12] The time it needs to equilibrate the energy distribution within the star through the random walk of photons.

[13] Gsponer's calculation [181] involves Newton's law and the disassembly time is given as $\tau_d = (r/c_S)\sqrt{M/m}$, where r is the radius of the fusion fuel sphere, c_S the speed of sound in the plasma, M the mass of the heavy coating material (the tamper), and m the mass of the fuel. Winterberg, on the other hand, estimates the disassembly time only with $\tau_d = r/c_S$. c_S has to be computed by a formula suitable for radiation-dominated plasma, and the two authors do it slightly differently.

the hot spot ignition mode a large sample of liquid fuel is set on fusion fire by just compressing and igniting a corner of it with the help of shock wave focusation. Subsequently, a fusion detonation front travels through the medium which is liquid or solid DD, DT, or solid ^{6}LiD. Inside the wave Winterberg estimates temperatures of 20 to 70 keV (1.7-8 $\times 10^8$ K, in the latter case in connection with a fourfold density increase behind the wave front, with respect to the initial liquid fuel) for the DT case and 300 keV (3.6×10^9 K) for DD. Whether the hot spot is a point from where the detonation wave spreads spherically or whether there is a cylindrical front originating from a hot spot in the shape of a rod, in both cases the flame surface has to grow, and the energy release per unit surface has to be high enough to sustain the burn in each next shell.

The alternative to the hot spot ignition is volume ignition, where the fuel temperature and pressure stay homogeneously distributed during compression and heating. In that case the energy balance has to be made up not for the flame front, but for the whole fuel load. The high-Z encapsulation being opaque to the plasma radiation helps in containing the energy release of the first occurring fusion reactions and supports *self-heating* in the fuel. This lowers the ignition temperatures. The thresholds for the fusion plasma to become self-sustaining can become as low as 2 keV (2.3×10^7 K) for DD and 1 keV for DT[181]. These low ignition temperatures are possible only if the compression is at least a factor of ten larger than in the hot spot mode. Compression ratios of 100 to 500 with respect to the initial liquid fuel ($n_0 = 5 \times 10^{22}\,\mathrm{cm}^{-3}$) are given. Historic hydrogen bombs can be assumed to have been laid out with sufficient margins so they could be ignited both ways, in the hot spot mode or via uniform fuel compression and heating, also because no failed tests are known [181, 517]. In the subsequent burn phase the temperature rises at least to 15 keV (1.7×10^8 K) and can reach up to 40 or 50 keV (5.8×10^8 K). In all these cases the plasma is radiation-dominated with $T_i \approx T_e \approx T_r$.

E.2.3 LCF - laser confinement fusion

LCF is actually another type of inertial confinement fusion where radiation-induced *ablation* of a pusher material is the driver of a concentric compression shock wave. The fusion fuel is encapsulated in a small spherical plastic pellet and frozen by cooling it to 18.6 K. The desired starting fuel distribution is an evenly thick layer of frozen DT coating the inner surface of the plastic pellet and the rest of the volume filled with DT gas at a pressure of several tens of bar. In the *direct drive* method laser light is used to heat the outside surface of the pellet directly. As the material at the surface is vaporised very fast and the only direction for the gas to move is away from the sphere, the resulting momentum stream propels the pellet material towards the centre of the sphere. (This process is called *ablation* and the resulting force, which can be imagined as a sort of rocket propulsion force, is called *ablation pressure*.) The resulting compression and heating of the pellet content produces a plasma where fusion reactions occur.

The largest LCF research facility at the moment is the National Ignition Facility (NIF) at Lawrence Livermore National Laboratory (LLNL) in Livermore (CA, USA). They have transitioned to a different method, *indirect drive*, which is based on the

hohlraum principle sketched out in figure E.2. The reason is that this method allows a more evenly distributed ablation pressure on the pellet surface and hence implosions with a better conserved symmetry.

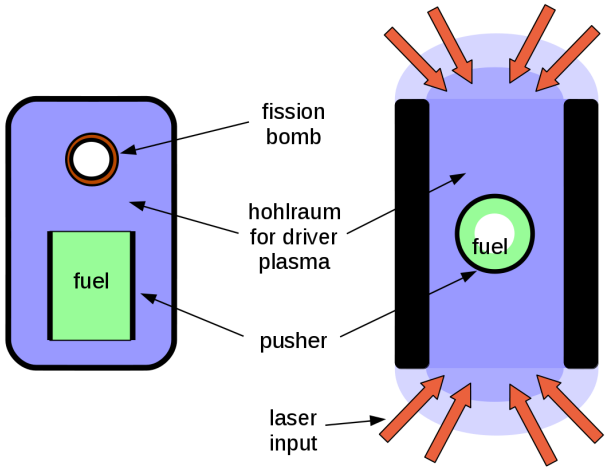


Figure E.2 The hohlraum principle of compression by surrounding plasma.
The sketch to the left shows a simplified schematic of the Teller-Ulam-Sakharov-Zel'dovich design of the hydrogen bomb and the one to the right the principle of the indirect drive of LCF. In both cases there is a thick mantle of high-Z material encasing a hohlraum volume painted in shaded blue. The hohlraum volume is to host a hot, radiation-dominated, thermalised plasma. This plasma is fed by a fission bomb (left) or by the laser radiation heating and ablating the inside of the gold cylinder and additionally creating a dense gas of X-ray photons (right). The purpose of the hohlraum is that the fusion fuel load gets completely immersed in plasma, and that the influences of the latter (heat, pressure, radiation) act equally from all sides. In both cases ablation of the outside surface of the pusher material in conjunction with the pressure drop from the hot hohlraum plasma down to the cooler fuel is the driving mechanism for the compression of the fusion fuel (light green).

The NIF hosts 192 laser beamlines producing infrared laser light ($\lambda = 1053\,\mathrm{nm}$) which is frequency-tripled to $\lambda = 351\,\mathrm{nm}$. A flash with an energy content of 1.9 MJ can be delivered within nanoseconds to the target. The engineering challenge connected with setting up the latest laser generation consisted in creating a large degree of flexibility in shaping the energy deposition rate over a pulse time of $\lambda = 20\,\mathrm{ns}$. After hitting the inside of the gold cylinder the energy deposited by the laser light creates a "nearly Planckian X-ray bath" [211]. The vaporisation of the fuel pellet outside surface creates an ablation pressure of $\sim 100\,\mathrm{Mbar}$ which leads to the shock compression of the hollow plastic pellet (initial radius about 1.1 mm and wall thickness 0.2 mm) and its fusion fuel content of about 200 µg. Inside the imploding sphere a fusion hot spot is created when the DT gas content is being smashed by the shock wave of denser DT from the formerly frozen layer. The results of the 2013 experiment campaign have just been published [211] indicating that the fusion energy yield from the hot spot is greater than the compression and heating energy delivered onto the whole mass of DT (which is however still only a fraction of the initial laser energy). The plasma hot spot reaches temperatures up to 4-5 keV (corresponding $46\text{-}58 \times 10^6\,\mathrm{K}$) at peak pressures between 1.26 and 1.52×10^{11} bar and produces a fusion neutron yield of about $\sim 5 \times 10^{15}$ at 14.1 MeV. A release of 17.3 kJ fusion energy was calculated for the strongest shot. The hot spot diameters measured by neutron and X-ray imagery are up to $\sim 100\,\mu\mathrm{m}$. Interestingly, the density disconti-

nuity representing the DT ice surface is present still when the hot spot flares up. At that time the density at the centre is estimated to be 34-50 $\mathrm{g\,cm^{-3}}$ while the value is 385-402 $\mathrm{g\,cm^{-3}}$ in the surrounding colder fuel. The higher value corresponds to a compression factor of greater than 1500. The peak fuel velocity estimation is about 300 $\mathrm{km\,s^{-1}}$.

In the context of LCF the comparison of hot spot and volume ignition mode has also been a point of discussions. The argumentation for the hot spot setup as described above and used at NIF is according to S. Pfalzner [347], that less energy is needed for acceleration and compression of denser and cooler fuel than for heating the whole fuel volume throughout and further compressing hot fuel. Adding more aspects to that point, she notes that, hotter electrons having a larger mean free path than colder ones transport ahead of the pusher ablation front or a fuel compression shock front leading to a preheating of the central fuel region and making its compression more costly. Furthermore, the pure hydrodynamics and thermodynamics of shock waves afford the result that shock compression requires a higher energy cost than adiabatic compression, but that replacing a strong shock with consecutive smaller shocks diminishes the difference [347].

Winterberg presents an interesting calculation example in [518] by asking: what radius does a sphere of liquid DT need to have, so that a single compression shock will be able to ignite a fusion burn front in the centre? His answer turns out to be 30 metres.[14] The current approach of NIF, the implosion of a shell of cold fuel smashing onto a small quantity of gas forming the igniting hot spot, can thus be seen as an attempt to take the scenario far beyond the boundary conditions of Winterberg's calculation example. The key is the increase of the compression factor, and the screws tweaked to do this are i. a. the right proportioning and spacing of fuel and pusher, the division of the fuel in solid and gas portions, and not least the shaping of the laser pulse in time [211, 347] aiming at putting the energy into accelerating the shell of dense fuel while not too much preheating the gas core.

E.2.4 Inertial-electrostatic confinement fusion (IECF)

IECF is the easiest and cheapest way to achieve fusion in steady-state plasma conditions. It has become the target of amateur and high school projects. The necessary components are a vacuum chamber and pumps, a special geometry of electrodes, a high-voltage supply, and a little bit of deuterium gas. Figure E.3 contains a schematic of a simple IECF reactor.

The basic design, sometimes also called a *fusor*, needs no magnetic, only electrostatic fields. It is a mixture between an old-fashioned TV tube and a gas discharge

[14] The compression ratio of a hydrodynamic shock can at most be $(\gamma+1)/(\gamma-1)$ where γ is the heat capacity ratio. For a mono-atomic gas with $\gamma = 5/3$ the compression ratio of a plane shock cannot be larger than 4. Two oncoming plane shocks will, upon rolling over each other, increase the density by a factor of 16. Going from plane to spherical geometry further adds a factor of 2, leading to a total density increase by the factor of 32. Unlike the compression ratio, which stays constant while the shock travels to the centre of the geometry, the temperature goes up. Assuming DT with the same density it has in liquid form at ambient pressure, but at a starting temperature of 10^4 K, the question is at what radius the incoming shock wave has to start so the centre volume with T over the threshold of 10^8 K will become large enough (radius $\geq$0.3 cm) so there is not only ignition, but burn propagation in the outwards direction.

lamp. One way to think of it is that a fusor is like a TV tube, just inverted and warped into spherical geometry. In a cathode ray tube electrons evaporate from a wire and are accelerated towards a ring-shaped anode at a positive potential. The electrons fly through the anode ring because they have been accelerated towards the centre of it. Before they reach the anode, the electrons are steadily getting faster due to the repulsive negatively loaded wires behind them and the attractive positively loaded wire ring in front. Once they are past the anode, both things lie behind and the two forces cancel out. This creates the electron beam that can be used for drawing a TV image on a fluorescing screen. The final speed and kinetic energy of the electrons is directly related to the voltage difference (potential drop) between cathode and anode. In a fusion device fast ions are needed, not electrons, so in a fusor positive ions are accelerated from the anode towards the cathode. Since the cathode is a wire grid ball positioned at the centre of a metal chamber serving as the anode, the geometry of the accelerating field is spherical instead of linear (see fig. E.3). The field-free space (field-free only in the absence of large densities of ions and electrons) reached by the ions past the cathode, is the space inside the grid ball. Here, ions incident from opposing directions can collide. If they had the possibility to pick up enough kinetic energy crossing the potential slope from the chamber wall to the cathode grid, then they can fuse after collisions. This is the simplified version of explaining a fusor.

A more realistic explanation has to account for the deformation of the electric potential by the distributions of ions and electrons and the possible particle trajectories in the modified potential wells. This allows to explain the main advantage of the concept, the minimised outflow of energy and particles from the active plasma region. The principle underlying the potential well deformation is illustrated in the right half of figure E.3. The negative cathode creates a potential well for positive ions. A gas atom becoming ionised inside the cathode sphere ends up with a relatively slow speed in some random direction. But an ion created in the outside volume gets accelerated strongly towards the central region, crosses it and bounces back on the potential slope on the other side. Such ions oscillate from side to side. The ensemble of moving ions in the chamber creates a probability density distribution with a peak in the middle. That ion probability density peak makes an upward dent in the potential well created by the anode and cathode. What is a potential peak for ions is a potential well for electrons. In a similar pattern the electron probability density can accumulate in the centre of that well and create another dent in return. Both things happen at the same time. While the ion accumulation is responsible for attracting electrons in the central region, the presence of the electrons helps the ions in not getting slowed down too much near the centre or be even bounced off before crossing it. One speaks of virtual anodes and cathodes. This illustrates again that plasma properties can often only be understood by the emergent effects of the collective motion patterns evolving interdependently for ions and electrons. The important thing is that ions can fly through the central region several times before being slowed down by collisions or captured by the cathode wires. The losses of particles and energy from the plasma are minimised by the device architecture and the establishing of beneficial plasma modes. This is why appreciable fusion rates become possible through a relatively low-tech lab setup. Already in the 1960s neutron

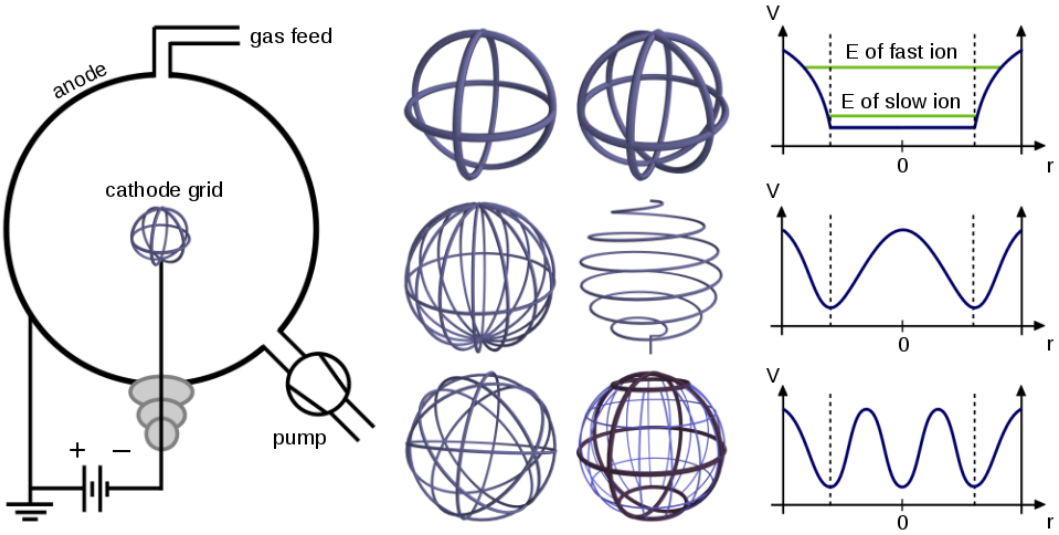


Figure E.3 Potential wells in IECF
A schematic of the fusor is shown on the left, several types of cathode grid geometries are depicted in the middle image, and the potential well seen by the ions is sketched on the right. The fusor is sometimes described as a circular particle accelerator: gas atoms are ionised by collision, then the positive ions feel the repulsion from the positively charged chamber wall and the attraction from the negatively charged cathode grid and are accelerated towards the centre. There, the ions collide and may undergo fusion depending on the types of colliding ions, their kinetic energies, and the collision angle. But there are several reasons why it is better to think of potential wells for ions and electrons instead of the slingshot image of the accelerator. They are (a) that the ions bounce back and forth several times, they oscillate across the centre or follow complicated (e. g. star-shaped) orbits, (b) the device is normally not used for pulsed shots but for creating a steady-state plasma, and (c) the interaction between the ion and electron populations can be understood easily by thinking of how one species modifies the electric potential well felt by the other. The upper plot on the right shows the electric potential existing in the chamber if it is empty or if there are only very few charged particles present. The location of the cathode grid is marked by the dotted vertical lines. An ion can "roll" from side to side in this potential like a marble in a salad bowl. Ions created far away from the cathode grid and near the chamber walls have the highest speed when they cross the centre region. Ions created by collision near or within the cathode grid reach only slow speeds. Many ions oscillating with an isotropic distribution of directions create a substantial probability density in the very centre at the point where all trajectories cross each other. Slow ions spending more time travelling across the centre have a higher impact building up the density peak as compared to fast ones. This accumulation of positive charge makes a dent in the potential well, this is shown in the middle plot of the right image. But for fusion reactions this central potential mound is very bad, it means slow collision speeds in the very centre (marbles are only fast at the bottom of the salad bowl). However, the accumulated positive charge attracts electrons at the same time: the central dent, seen upside down, is a potential well for electrons. Inside their potential well electrons can follow the same pattern as the ions in their well. Electrons accumulating in the centre region of their well will make an upward dent in the potential in turn. This is shown in the bottom plot. The nested wells have been termed *Poissors* by Philo Farnsworth, one of the inventors of the fusor [300]. Experimental data can be gathered representing a three-peak structure when scanning D-D fusion rates via direction-sensitive detection of fusion neutrons [182], thus confirming the triple ion trap structure as plotted in that last image. What prevents a further continuation of the chain of nested wells is the spread in particle velocity vectors and angular momentum [300].

production rates of up to 10^{10} neutrons per second could be achieved [208].

Variations of the IEC setup

The basic concepts of IEC plasma devices can be expanded to a variety of operation modes and application scenarios. Figure E.4 depicts various plasma modes that can be achieved with variations in the gas pressure. More modifications of the described simple setup led to the following designs and devices [300]:

- Cathode and anode can be swapped. Replacing the cathode grid in the centre by an anode grid turns an ion-injected into an electron-injected IEC. In both types the probability densities of both ions and electrons accumulate in the centre.
- A simple method for not having to rely on background ionisation is the installation of a filament. It has to be placed near the chamber wall because only ions created far away from the cathode end up gaining a lot of kinetic energy.
- Multi-grid devices: the anode can be made up by spherical grids instead of the chamber wall; ion creation can then be achieved with RF fields between the outer grids. Lower pressures become possible. Another advantage is more flexibility for the shape of the vacuum chamber.
- Instead of relying purely on ionisation in the main chamber, either by the grids or by an additional filament, external ion cannons can also be used. This reduces the energy carried away from the plasma by electrons.
- The normally independent back and forth oscillations of the many ions across the central potential well can be tuned to be in phase through radio frequency excitation and with the help of additional electron injection for shaping the potential well in the right way. This has been discovered by Park et al. [341]. Due to the periodic concentric clashes of the ion cloud this is the most interesting comparison of IEC methods to SF (see below).
- The multiplication of the star mode in figure E.4 with the principle of synchronised clashes in a multi-grid device leads to the improved IECF reactor concept investigated with simulations in [294]. At low background pressure pairs of ion bunches oscillate along crossed beam paths and synchronously clash in the centre of several spherical grids. The simulations indicate that a mechanism of self-organisation could be used for concentrating and synchronising the ion bunches. Experimental confirmations of the principle seemingly have not yet been published, but the concept is nevertheless interesting enough to be included in the discussion due to the implications on how to deal with the basic thermodynamic differences between beam-like fusion reactors and systems in thermodynamic equilibrium. The key aspects are explained below.
- An IEC device operated in the jet mode as shown in figure E.4 is envisioned as propulsion engine for space missions [239, 300]. It could be open to the surrounding space with no need for a heavy vacuum chamber. Feeding on

electric power from solar panels, it would be classified as electrically driven thruster. A fusion power input to the energy balance is not necessarily part of a useful thruster design. Heavy non-fusing nuclei can be favoured for momentum generation.

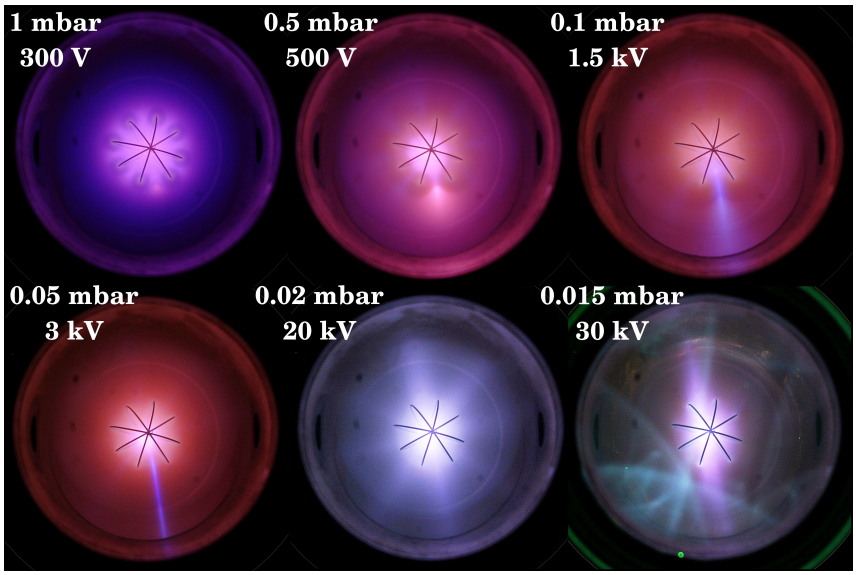


Figure E.4 IEC plasma modes
These photographies show discharge plasmas in an IEC chamber filled with air. Variations in gas pressure and driving voltage lead to very different plasma states. From top left to bottom right the first four snapshots show how the transition from the *central spot mode* via the *spray* or *halo mode* to the *jet mode* can be controlled by decreasing the pressure. Using a constant series resistor, the voltage across the chamber electrode grows as the conductivity of the plasma decreases. Because of the existence of the jet mode researchers investigate the use of IEC plasma as propulsion method for space missions [239, 300]. The location of the jet can be pre-determined by introducing asymmetries in the cathode grid like cutting out wire pieces [239]. The fifth image shows the operation in *star mode*. In that mode a symmetric array of brightly shining light beams is visible. The evenly balanced rays represent accumulated trajectories of ions oscillating from side to side. This mode is suitable for achieving high fusion reaction rates [300] when using deuterium gas instead of air. The greenish light patterns in the last image arise from fluorescence in the glass. These photographs have been kindly provided by T. Rapp [376].

Periodic ion cloud contractions excited in an IEC device

Can the oscillating motion of different ions be tuned to be in phase? This replaces the stationary plasma state where the times when different ions cross the centre region are not coupled by a state of collective motion of the particle species. The existence of such a radial collective plasma oscillation has been proven experimentally by Park et al. [341] through excitation with radio waves. They call it a *periodically oscillating plasma sphere (POPS)*. As the frequency is independent from the ions' motion amplitude, they assume a parabolic potential well shaped by the underlying electron gas. An additional electron injection is a key part of the experimental setup, it supplies the right background electron density. The periodic concentric clash of ions reminds of sonoluminescence, and there is indeed a reason: the termodynamics become more similar. Thinking of harvesting fusion energy, the interesting difference between the POPS mode and conventional stationary IEC plasma is this: In the star

mode newly incoming ions make a few passes through the cathode centre at high speed before they get slowed down (thermalised) by collisions. According to Park the "high energy cost of maintaining a beamlike ion energy distribution makes it difficult to produce net fusion power and is considered to be a crucial obstacle facing IEC based fusion energy devices" [341]. In steady-state IEC plasma thermalisation of ions means their slowing down and the randomisation of their motion directions. But in the POPS there is a local thermal equilibrium in the ion population anywhere anytime, outlier ions with odd motion patterns are pulled back into the collective pattern. Going from *beam-like* plasmas to conditions at *thermal equilibrium* when achieving a fusion energy break-even means avoiding a substantial fight with the laws of thermodynamics.

Self-organisation among oscillating ion bunches

In his doctoral thesis T. McGuire investigated a type of IECF reactor in which electric fields lead to the stabilisation of oscillating ion bunches with beam-like velocity distributions. The concept is described as a principally viable way towards a break-even fusion reactor, although with narrow limits in terms of fusion power densities. This seems to be a brazen contradiction of Piel's [351] argument[15] that fusion-capable beam-like setups must remain net energy consumers. Perhaps the seeming contradiction can be appeased by noting that McGuire's setup can be seen as a link bridging the large gap between linear single-beam devices and the thermodynamically favourable POPS approach mentioned above.

With theoretical calculations and plasma particle simulations McGuire investigated how IEC devices can be made more efficient in terms of fusion rate and neutron production. He put the main focus on enhancing the ion lifetime. In a conventional fusor with a single cathode grid ions live so short that they can only pass the centre 1-10 times until they are thermalised or hit a cathode wire. A multi-grid structure holds the potential for enhancing the ion lifetime. Operating a multi-grid IEC chamber in star mode is only possible if the grid layers are aligned with each other so that straight flight paths exist for ions crossing from side to side. The multi-grid structure can be used to fulfil three additional purposes for improving the properties of the ion trap: (a) The layers of wire grids can make the electric field patterns more symmetric and in particular reduce the impact of the asymmetry arising from the electric feed lines to the grids. (b) Relaxing the notion that the electrode grids must be in the form of "wire grids" and making the transition to metal spheres with precisely machined circular holes, the distances and potential drops from metal sheet to metal sheet in conjunction with the orifice radii can be used to engineer focusing electrostatic lenses[16] along the ion flight paths. (c) The electric potential normally drops from grid to grid in the inwards direction. Reversing the direction from the second innermost to the innermost grid has a decelerating effect on ions but creates a potential well for electrons and minimises their losses to the anode wall.

[15] see page 226

[16] weblinks explaining charged particle optics: (i) `http://en.wikipedia.org/wiki/Electrostatic_lens`, (ii) `http://encyclopedia2.thefreedictionary.com/electron+lenses`, (iii) `http://engineering.siu.edu/frictioncenter/cafs-courses/surface-contact-mechanics/lecture-4.php`

The increased electron density in the core minimises the defocusing effect of the elevated central positive charge accumulation created by the ion beams themselves. All these tweaks and innovations should allow running the IEC chamber with a lowered background pressure (less thermalising collisions with background gas atoms, lower input rates of ions and electrons needed from ionisation) and prolong the lifetime of oscillating ions so they can cross the central collision region many times at high kinetic energy.

McGuire analysed the movements of the ions in the IEC chamber with numerical models, and the most interesting finding comes from that part of the work. Particle-in-cell[17] simulations showed that under certain conditions the Coulomb interaction among ions can lead to synchronisation among the oscillating ions. The synchronisation occurs not only among ions of one single ray of the plasma in star mode, it also happens between the different rays. On each straight pathway two opposing ion bunches will form. In a symmetric pattern they return at the two path ends at the same time and run over each other head-on in the centre at the next moment.[18] The bunch formation is understood as an instability mechanism with a limit amplitude spreading from ray to ray. However, the simulations show that this collective phenomenon emerges only if the ion density is in the right range and if the ion lifetime is long enough. Ions running ahead or lagging behind are pulled back into the bunch. This means through the Coulomb interaction kinetic energy is transferred back into ions which have become slower than the rest of the bunch. McGuire writes: "The synchronisation mechanism provides a concrete example of how non-Maxwellian, non-neutral plasmas can be maintained by using the very forces that usually thermalise it." This would be the crucial thermodynamic game-changer if it should become possible to exploit the effect for a fusion reactor because it would reduce[19] the importance of Piel's argument. However, an important limitation is

[17] In a particle-in-cell simulation the electromagnetic fields are represented on a grid. The Maxwell equations are solved on that grid. The movement of charged particles through the volume is simulated with standard time-integration schemes. The summed charged particles present in a grid cell modify the electromagnetic field. Field updates follow particle position and velocity updates and vice versa. Inter-particle interactions are accounted for only indirectly via the fields. Often macro-particles are used, whereby n electrons or ions are represented by larger particles of the same charge-to-mass ratio.

[18] Why does the Coulomb repulsion lead to the formation of coagulated ion bunches? Why do the ions not stay evenly distributed on their linear oscillation track? This can easily be understood by considering two simple thoughts. The first thought experiment is to imagine two ion bunches going into the same direction at the same velocity, one bunch following the other. Repulsion between the two bunches accelerates the leading one and slows the lagging bunch. In the closed oscillatory movement along one ray of the star mode, the front-running bunch turns into the lagging one as soon as it is more than 180° ahead. From that moment it is the one being slowed down. The second thought experiment is needed to explain why one single ion prefers to be part of one of the two bunches. Why is a front-running ion not accelerated and pushed further ahead and a lagging ion not further slowed and expelled from the back end? It can be explained by considering an ion at the centre of mass position inside one bunch: the repelling forces from the ions of the own bunch annihilate each other and what remains is only the repulsive force from the other bunch. Therefore, the bunch centres can be seen as attractors for single ions. With this bunching and synchronisation mechanism there is even coupling between neighbouring rays, but the angle is important: the smaller the angle between two rays, the larger are the regions where the two tracks run in close proximity.

[19] Piel's argument will not be completely invalidated because when two ions from two bunches collide (e. g. head-on) then there is still the possibility that they undergo only elastic scattering

already stressed by McGuire, namely that at too high ion densities the repulsion between the ions prevents bunching and synchronisation.

Adding magnetic fields to the IEC setup

The addition of magnetic fields bridges part of the way towards magnetic confinement setups. In the two approaches outlined below the magnetic field is used to keep electrons confined so they form a virtual electrode used for confining the ions electrostatically in return. The advantages of grid-less designs with mainly virtual electrodes are that there is no danger of melting grids and the avoidance of losses due to ions colliding with the grid.

- The polywell is a device in which multiple coils create a magnetic field with cusps for trapping the plasma. It is explained in figure E.5. Its problem are electron losses through escape paths following magnetic field lines.
- In a Penning or Penning-Malmberg trap the space between the electrodes is very small and offers only millimetre or centimetre volumes for plasma trapping. The whole electrode construction is placed in a uniform $\vec{B}$-field. The advantages of that type of setup are the minimised electron losses and the high fusion power densities. The disadvantage is that the device size cannot be increased so that upscaling is only possible via large arrays of small exemplars ("the Penning trap must remain a one-Debye-length machine" [78]). The basic principles are explained in figures E.6 and E.7.

E.2.5 Magnetised target fusion (MTF)

In order to continue the description of plasma confinement mechanisms in the order of increasing influence of magnetic fields and decreasing influence of inertial forces, techniques of *magnetised target fusion (MTF)* shall be described next, before finishing with purely magnetic confinement (MCF). According to [521], the concept of MTF stems from Russian research of the 1970s [127]. The main idea is to leverage the physics of electromagnetism when compressing plasma with metal liners for keeping the plasma itself out of touch of the metal surround with the goal of minimising energy losses through heat transfer.

Fast liner compression experiments at LANL

Lenz's law says that currents induced by changes of magnetic fields flow in such a direction that their own contribution to the magnetic field opposes the initial change. One implication is that with a coil, a hollow metal cylinder and some explosive material one can create magnetic fields of several hundred tesla. The trick is to use the explosive to implode the metal cylinder very quickly after having created a magnetic field inside the cylinder going in the axial direction. The explosion-driven

without fusing. However, a good fraction of scattered trajectories can be recovered because ions can continue on other oscillation tracks (other rays of the star mode) after being deflected by a large angle or be absorbed back into the own bunch after small trajectory perturbations.

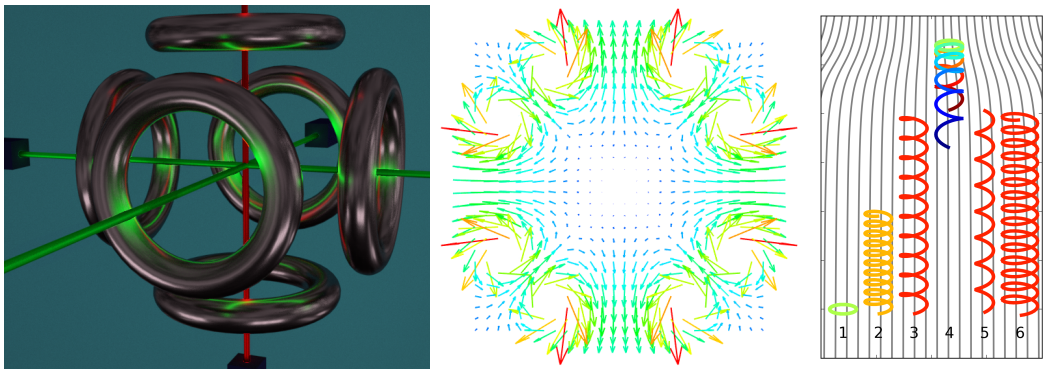


Figure E.5 The polywell

In a polywell the ions are electrostatically confined by a cloud of electrons forming a virtual cathode. The electrons however are magnetically confined in the field created by six coils arranged in a cube-like form. Such a coil arrangement is illustrated in the first image where the location of electron and ion insertion beams is indicated by the green and red lines. Sketching the magnetic field in a plane cutting the cube and four of the coils in half yields the second image. Two opposing coils create a field with $\vec{B} = \vec{0}$ in the central plane. The field created by the array of six coils is zero in the centre point, weak in a region around it, it gets stronger towards the coils, and field lines form cusp structures in any direction. Why these cusp structures help to form an electron cage is illustrated in the third image.

Because of the Lorenz force $F = q\vec{v} \times \vec{B}$ electrons move along circle lines or helical paths in a magnetic field. A second concept necessary to understand the motion of charged particles in magnetic fields is the concept of a magnetic mirror. Both concepts are shown in the third image where the colour scale codes for the particles' vertical velocity component. Trajectories 1, 2, and 3 represent electrons with the same radial velocity but different vertical speeds. They extend only in the area where the magnetic field is homogeneous as indicated by the parallel field lines. The 4^{th} trajectory starts with positive vertical speed coded in red and enters the region of narrowing field line distances symbolising a growing $|\vec{B}|$. As a consequence the electron slows down its vertical upward motion and turns downward to follow the blue part of the helical trajectory. In reflecting the upward into a downward motion the inhomogeneous magnetic field is said to act as a *magnetic mirror*. Placing an equivalent magnetic bottleneck on the other side will create a *magnetic bottle*. Trajectories 5 and 6 depict particle tracks with different helical radii. This can be the sign of either different radial velocities or different mass-to-charge ratios. In the case of identical particle type and within the symbolism of this plot where the colour codes for vertical speed and thus a helix oscillation time scale it can however only mean a radial velocity difference.

Magnetic mirrors in the form of field line bottlenecks are never absolutely tight. Electrons with sufficient forward momentum are able to overcome the barrier, they can continue their corkscrew path all the way through until the field lines widen again. The polywell features such bottleneck pathways in any direction. Electrons are inserted into the cube with electron guns and ions with ion cannons. To use the magnetic field to trap the ions directly would be harder because they have so much more mass and inertia per charge leading to too wide spiral diameters. The electrostatic confinement of a minority of ions by a majority of electrons in a non-neutral plasma is the main working principle of the polywell. The main hurdle preventing fusion energy break-even with the polywell technique is the leakage of electrons out of their magnetic mirror cage [78]. On the one hand any limit of electron density translates directly into a limit of ion density. On the other hand the necessary equilibration of the losses with incident particle beams costs energy. An interesting feature of the polywell technique is that the plasma can be made so dense that surface currents become capable of creating a field-free inner plasma region. Newer research aims at exploiting the formation of that skin of surface currents for reducing electron leakage [342]. When the surface currents are freeing the inner region of the plasma from the magnetic field, it means that they create a counter field and push the $\vec{B}$-field lines outwards. This can be imagined as a magnetic counter-pressure of the plasma equalling the pressure of the external magnetic field. Such plasmas are called high-β, whereby $\beta = P_{\text{plasma}}/(B^2/2\mu_0)$ is the ratio between plasma pressure and magnetic pressure [342]. The existence of an extended field-free inner plasma region means $\beta = 1$. For comparison: MCF machines like a tokamak work with $\beta << 1$, e. g. $\beta \approx 0.03$ is envisioned for ITER [122].

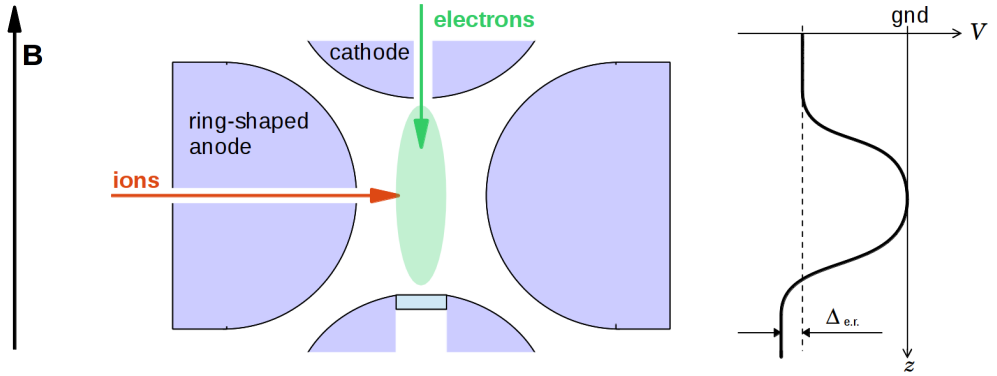


Figure E.6 The Penning trap
The sketch shows a cross section of the geometry of a Penning trap. It is a radially symmetric arrangement of two cathodes opposing each other and pointing at the centre of a metal ring constituting the anode. The green arrow indicates an incident electron beam coinciding with the rotation axis. If the electrons are not desired to hit the lower cathode and be absorbed by it, this can be prevented by using electrons of low kinetic energy and by decreasing the electric potential of the lower cathode (by a $\Delta_{\mathrm{e.r.}}$) so it can function as an electron reflector. Without any magnetic field the electron density in the trap volume would stay low as electrons would be quickly attracted and absorbed by the anode. But placing the whole setup within a magnetic coil and immersing it in a uniform $\vec{B}$-field aligned with the rotation axis changes the situation. Now the electrons cannot reach the anode so easily any more because the $\vec{B}$-field forces them onto circular orbits or spiralling tracks. On these tracks electrons move up and down in the potential well plotted as V over the vertical coordinate z on the right. Similarly as in the polywell, the trapping and accumulation of a dense electron gas by a magnetic field allows the formation of a virtual cathode which can be exploited for the inertial-electrostatic confinement of an ion plasma. Ions can be inserted into the Penning trap through a borehole in the anode ring. The depth of the potential well formed by the virtual cathode determines the maximum collision speed of the ions. Diagnostic signals can be gathered by placing a metallised lens at the tip of the reflector cathode.

fast motion of the metal wall along the radial direction and thus orthogonal to $\vec{B}$ induces large circular currents in the metal due to the Lorentz force. Of course, unless it is in the superconducting state, such circular currents in the metal would quickly decay through ohmic losses. But during an explosive-driven implosion there is not enough time for that, the ohmic losses are negligible, and thus the magnetic flux enclosed in the conducting loop is being conserved during the contraction. The consequence is a diverging flux density $|\vec{B}|$. The creation of the circular currents with their potential of heating the metal through ohmic reminds us that work needs to be done when moving conducting circuits through magnetic fields in certain ways. Eddy current brakes would be another application of the same principle.

Researchers at Los Alamos National Lab (LANL) make use of these effects just the other way round: they use huge current pulses to build up a large magnetic field. The quickly growing $\vec{B}$-field performs work on a metal cylinder and forces it into implosion. When inserting magnetised plasma into the cylinder beforehand, the implosion can be used to compress and at the same time heat the plasma. Figure E.8 shows the experimental setup providing The first part, the imploding metal liner.

Figures E.9 and E.11 show how to get the second ingredient for such a type of fusion experiment: self-confining magnetised plasma. Two forms of it are explained in the pictures, so-called FRCs and spheromaks.[20] Understanding how such plasma

[20]FRCs and spheromaks are two different ways of creating magnetised plasma in the form of a so-called *compact torus*. The field geometries are similar to the ones used in MTF reactors (see

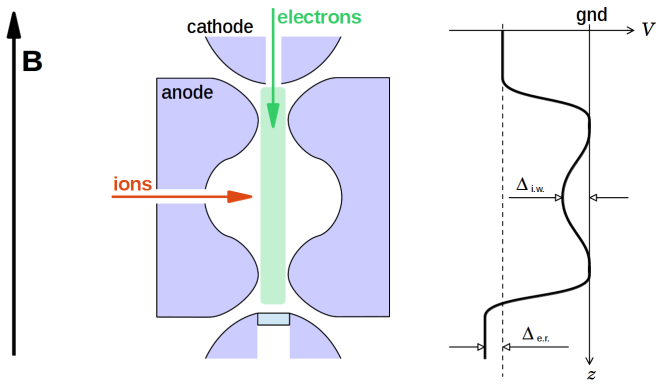


Figure E.7 The Penning fusion experiment
This drawing shows a modified Penning trap adopting features of the electric field from a so-called Malmberg trap. It offers a better axial confinement for the ions and has been proposed by a group around G. H. Miley (University of Illinois at Urbana-Champaign) and D. C. Barnes (LANL) for a *Penning Fusion eXperiment (PFX)* [78, 402]. Schauer et al. [402] explain the advantage of the modified design by reminding that the electrons are not thermalised and by pointing to the concept of image charges. If the electrons have a beam-like non-thermalised velocity distribution it also means they are not bunched in shallow local potential wells and do not equilibrate charge density perturbations of the plasma. Instead, they bounce back and forth between the two cathodes and create a band of rather uniform charge density near the central axis symbolised by the green shading. By their presence, these electrons induce an additional electric field component in the volume between the electrodes and modify the electric potential V. The important features of the field and potential can be explained with the concept of image charges and the underlying reality of surface charges in the metal parts. Image charges are an imaginary concept, a tool for simplifying construction and analysis of electric fields. The decisive reality is that free charge carriers in a metal equilibrate all potential gradients, including gradients along a surface, so that all $\vec{E}$-field lines end up meeting conducting surfaces orthogonally. While the cavity in the anode can be seen as a Faraday cup of which the inner volume is supposed to be field-free (in terms of electrostatics), the electron density along the central axis changes the situation. If an electron is positioned off-centre in a Faraday cup, the field lines are denser towards the closer wall. So the electron is not force-free, the potential energy of the system is lowered when the electron approaches the wall, and work is done on the electron being attracted by and accelerated towards its image charge. This is why the cavity in the anode has the effect of creating an energy barrier for the electrons because in the cavity's centre where they are further away from any wall the electrons have a higher potential energy than in the bottlenecks above and below it where the wall is closer. A secondary effect is that the energy barrier slows the beam of electrons and increases their density. The condensed background of electrons, functioning as a virtual cathode, even further raises the potential energy of any single electron under consideration. This peak of the electric potential V into the negative direction on the height of the cavity is illustrated on the plot all the way to the right where V along the central axis is plotted over the space coordinate z. If it is a barrier to electrons, then it appears as a potential well to ions. This ion well depth is indicated as $\Delta_{\text{i.w.}}$. One could however ask why the image charges of the ions have been totally left out of the discussion. Why does the cavity in the metal block not equally pose a barrier to the ions? The answer is that there is just one single well-determined function $V(\vec{x})$ and one single vector field $\vec{E}(\vec{x})$, both being the sum (linear superposition) of the contributions from all the charges in the electrodes and the plasma. If ions are the minority and electrons by far the majority in the non-neutral plasma then electrons can be imagined as forcing their scenario upon the ions. The PFX ion trap can only function and provide a deep well for fast ion collisions if the plasma is far from neutral.

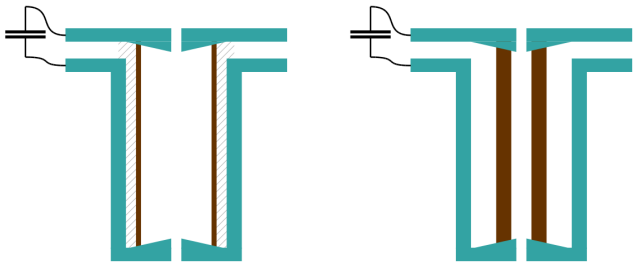


Figure E.8 Magnetically driven fast liner implosion
Researchers at LANL use huge current pulses to bring metal cylinders to a magnetically driven implosion. In the exemplary experimental setup [112] depicted above the cylinder to be imploded is shown in brown. It is made of aluminium. The surrounding metal construction shown in blue is made of a harder metal. The current source is indicated by the small capacitor symbol on the left. In reality it consists of large banks of many capacitors connected in parallel and equipped with special fast-acting switches. When these switches are closed and form the circuit symbolically depicted above, then the sudden unloading of the capacitor banks creates a large current pulse which in turn leads to the fast growth of the intended magnetic fields. The relevant volume portion is hatched in the sketch. The current flowing around that volume creates a magnetic field with circular field lines in between the metal cylinders. The quickly growing field pushes the sturdy cylinder wall outward and the soft one inward. Only the soft one gives in. So the field inside the current loop is used to crush the aluminium cylinder. The sketch shows that conical end caps of the sturdy cylinder are one possible measure ensuring that the compression of the inner cylinder does not rupture the electric circuit. The problem with this construction is that access to the inner volume is restricted by the small outlets. In [112] Degnan et al. also present an alternative solution where only the central part of the softy cylinder is radially contracted and no sliding bearings are needed. This means the whole diameter of the inner volume can be accessed and used for plasma insertion, as the purpose of the whole setup is to provide a fast imploding metal liner for compressing magnetised plasma. [451]

structures can be formed and what keeps them together requires the consideration of several more application cases of the Lenz rule. Looking at these cases one could say that plasma can act like nails connecting two wooden pieces: it can clamp the two worlds of matter and electromagnetic fields together so that they cannot move freely against each other any more.

Having understood the two phenomena that there are stable and self-confining plasma structures and that sideways moving magnetic field lines can transmit forces, it becomes clear why compressing magnetised instead of normal plasma with a metal liner has a huge advantage: greatly minimised energy losses due to diffusive heat transport. When a spheromak or FRC is compressed by a collapsing metal liner not much heat is conducted away from the plasma via the surrounding metal because the metal is not very much in contact with the plasma. The transmission of mechanical forces works via the magnetic field. As the diameter of the metal-lined cavity shrinks, the induced currents in the liner diverge. The magnetic field arising from the currents in the liner can thus keep up with the rising mechanical and magnetic pressure of the compressed plasma structure.

This has good and bad consequences when thinking about how to modify and scale up experimental setups for harvesting fusion energy. First the bad side: either there has to be a factory for manufacturing metal bottles in front of the reactor where they have to be inserted, connected to the vacuum system, and imploded in quick succession [419] or one has to come up with a method of nicely imploding cavities in liquid metal. On the good side it has to be noted that MTF approaches promise

below). The only difference is that there is no central pole of structural materials enclosed. For this reason the term *compact torus* is used to distinguish from the torus geometries common in MCF.

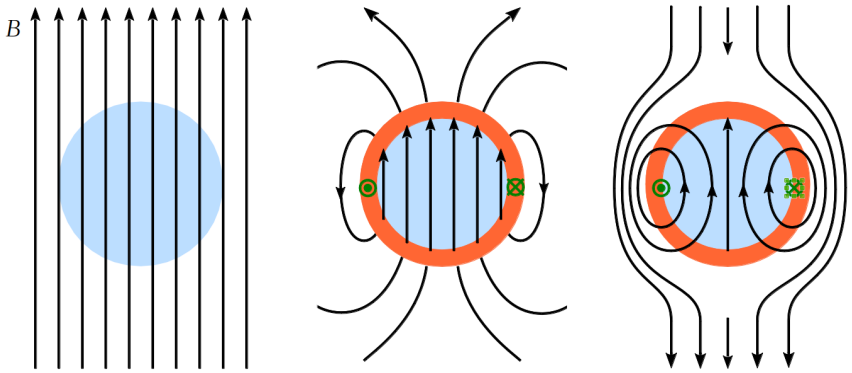


Figure E.9 Plasma in field-reversed configuration (FRC)
In a similar way as the magnetic flux can be conserved inside a conducting loop, a magnetic field pattern can be "frozen into" a plasma. This is symbolically depicted in the snapshot sequence above. On the left, the blue region indicates a plasma created in a volume where a uniform magnetic field had been established beforehand. As plasma is a conducting medium, any change of the surrounding B-field will induce electric currents inside it. By Lenz' rule the additional magnetic fields created by these currents oppose the initial external change. This is why in a dense plasma skin-like current layers can arise. Since the currents within the skin have a compensating effect they act like a shield. The field change is not communicated further inwards, the plasma has the effect that the magnetic field is "frozen in". The middle image shows the situation with the frozen field in the core region of the plasma after the external B-field has been switched off. The shielding layer of currents is indicated by the orange colour. The green vectors show the general direction of the compensating currents, they follow horizontal circles around the plasma core (toroidal currents). It might be perceived as confusing that this image with "switched off" external field does not show any field-free regions. This is because only a close-up is shown. The magnetic field arising from the toroidal currents decays far away from the structure. The last image shows the situation after the external field has been re-established but with reversed direction. This leads to a compression of the closed field line loops around the toroidal currents. The resulting rotationally symmetric plasma structure is a so-called *field-reversed configuration (FRC)*. FRCs are a preferred study object of plasma scientists because of their inherent stability giving them liefetimes of *self-confinement* in the millisecond range. The stability arises from the energy stored in the magnetic field. When the toroidal currents weaken due to thermal losses, then portions of the energy from the decaying magnetic field get consumed for feeding the weakening currents. The thermal losses are the bottleneck through which the whole field energy has to go. This valve determines the stability timescale of the self-confining plasma structure. The FRC can be called a self-confining plasma structure because as long as the fields are strong enough charged particles following field lines in helical trajectories have no escape paths. Researchers at LANL are working on using the stability and extended lifetime of FRCs and the possibility to push plasma with magnetic forces for experiments where FRCs are inserted into a metal liner as depicted in figure E.8 before its implosion.

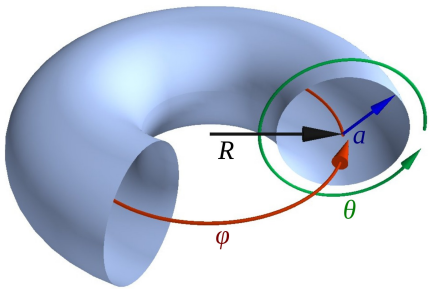


Figure E.10 Toroidal and poloidal coordinates
For describing the current and field line directions of plasma structures in the shape of a *torus* (doughnut) the terms *toroidal* and *poloidal* are very useful. In the above coordinate system φ is the toroidal direction and θ indicates the poloidal direction or coordinate.

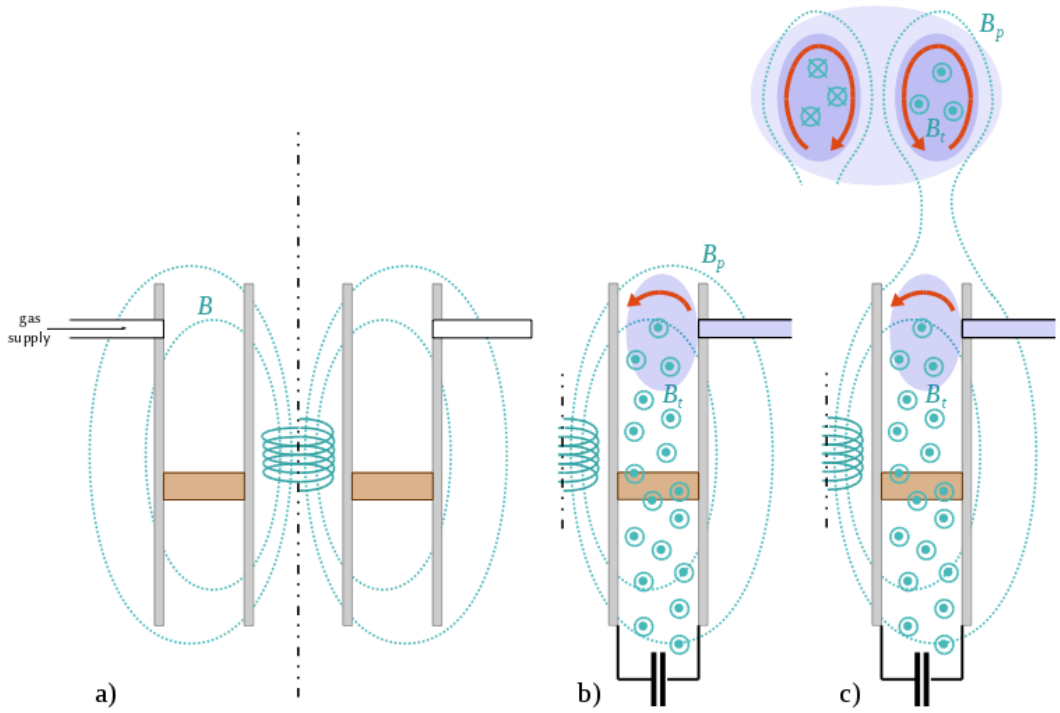


Figure E.11 Spheromak formation

Another important plasma structure with a degree of inherent stability is the *spheromak*. It is almost the same as an FRC except that $\vec{B}$ has non-zero toroidal components. The symbolic sketches above show the steps of creating a spheromak in an exemplary lab setup according to [34]. Toroidal and poloidal components of the magnetic field are labelled B_t and B_p. The key is an electric discharge between two concentric electrode rings whereby the gap is overlaid by the inhomogeneous magnetic field created above a coil. This rotationally symmetric setup is shown in image (a). The central coil and its B-field are both depicted in green. The remaining images only show the right half of the symmetric cross section. The spheromak formation is initiated when gas is puffed into the vacuum and a capacitor bank is discharged through the gas puff. The discharge current is shown as red arrow in image (b); it follows the magnetic field lines. The current looping from the capacitors through electrodes and gas back into the capacitors feeds a magnetic field with toroidal field lines lying orthogonal to the drawing plane. The mechanical pressure arising from the field growth (the same force that compresses the metal cylinder in figure E.8) pushes in the outwards direction anywhere along the current loop but the discharge plasma is the only link weak enough to be set in motion. When the gas begins to move upwards the special properties of the plasma come into play. The plasma carries the frozen-in magnetic field lines away and expands them like elastic rubber bands. The paths of the discharge current are taken along as they move along the magnetic field lines. When the field lines and currents rupture and reconnect a spheromak has been created. Conserving its momentum it moves away from its creation site as far as space and lifetime permit. The key difference with respect to FRCs is the poloidal component of the magnetic field. It will always arise if the plasma creation is based upon the discharge across a radial electrode gap. The sum of the toroidal and poloidal B-fields leads to helical field lines wrapped around the torus structure.

to be still low-tech in comparison with laser-driven ICF or MCF[21]. MTF combines the physics from both sides and some researchers argue that there may be a cost minimum in between the extremes of ICF and MCF. Discerning between liquid and solid liner MTF, besides avoiding a cartridge factory the liquid metal approach has other important inherent advantages: if all of the radiation is dumped in the liquid metal and turned into heat there then there is no need to develop neutron-resistant construction materials, liquid metals are ideal for transporting and exchanging heat, and the breeding of fuel based on lithium could also take place directly in the liner material.

Two MTF concepts shall be described in more detail. The MAGO experiments performed at the All-Russian Scientific Research Institute of Experimental Physics (VNIIEF) show that in conjunction with employing some tricks of electricity and magnetism intelligently MTF opens up a way to achieve thermonuclear fusion in an academic setting with relatively high energy densities in a 10 cm volume. The other one is the liquid liner approach taken by General Fusion®, a Canadian company founded not so long ago in the belief that with MTF there are no more unsolved fundamental physical problems remaining any longer between today's demonstration experiments and commercial electricity production in the near future.

MAGO experiments at VNIIEF

MAGO is the abbreviation for "магнитное обжатие"[22] which is Russian for "magnetic compression" or "magnetic implosion" [154, 155] and the term coined by researchers of the All-Russian Scientific Research Institute of Experimental Physics (VNIIEF) for a single-shot destructive small-scale experiment (chamber diameter 16 cm) achieving thermonuclear fusion (but not ignition) and relying on extremely strong current pulses. Two versions of the experiment, with and without imploding liner compression of the generated magnetised plasma, are explained in figures E.12 & E.13. Not only for plasma formation, jet acceleration, heating, and compression but also for the electric powering the physical phenomena are leveraged which have been described above in figures E.8, E.9, & E.11.

The basic principle of the MAGO experiment is the extreme acceleration of current-bearing plasma by using magnetic pressure to drive it through a narrow nozzle. The jet impacting on gas at rest is the ionisation and heating process for the gas volume behind the nozzle. Neutron measurements show that a deuterium-tritium gas mixture can be brought well into the fusion regime with this mechanism. Direction-sensitive neutron detection reveals that fusion reactions occur at a low rate inside the jet where fast and slow ions collide in friction layers or after particle collisions with the chamber wall. The highest fusion rates are however recorded in the central regions of the second chamber. Around 4×10^{13} fusion neutrons are generated per shot.

The plasma creation starts in the nozzle region. Not only the electric potential distribution along the metal surfaces and the potential gradients in the inner volume are the determinants of that location. An important factor is also given by the

[21] *magnetic confinement fusion*, see below

[22] possible transcriptoin: "magnitnoye obzhatiye"

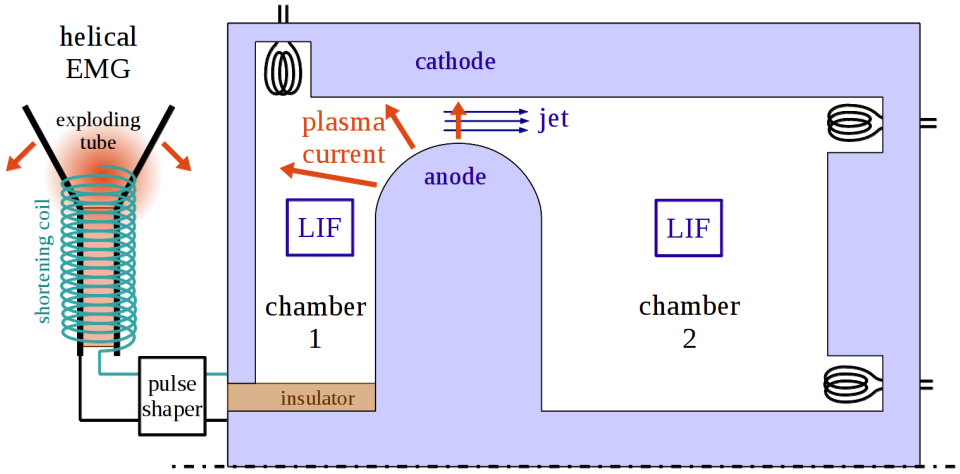


Figure E.12 EMG-driven MAGO chamber
In the MAGO experiment by Garanin et al. [154, 155, 270] magnetised plasma in a deuterium-tritium gas mixture is heated with a supersonic jet driven by magnetic forces to fusion conditions. Extremely strong electric current pulses are needed and can be supplied with explosive magnetic generators (EMG). The first such power source is depicted on the left: a capacitor bank discharge pumps a large current through a coil, and when the current peaks and a large magnetic field is established, then a metal tube placed inside the coil is expanded by a detonation front running through the explosive-filled tube. This short circuits one winding after the other while the magnetic field decays and releases its stored energy into an ever shrinking number of windings. The result is a large current pulse in the range of several megaampère. With a pulse shaping unit the pulse is divided into a slowly rising precursor and a fast rising main pulse. The power source and the chamber layout are dimensioned such that the precursor (max. current 2.7 MA) establishes a magnetic field throughout the chamber but creates no gas discharge yet. The signals from the three inductive field probes positioned in wall grooves will be in line. When the fast-rising main current pulse arrives, the current finds a shorter way directly through the gas which becomes ionised. A plasma current is established bridging both the narrow nozzle ring connecting the two subvolumes and the broader chamber 1 itself. The external current increases to 7.5 MA within 2.5 µs. The induction probe signals will now differ hinting to an important development: the new shortened current loop creates a strong B-field particularly in chamber 1 leading to a pressure difference and the formation of a jet from left to right. The magnetism-induced pressure difference is strong enough to turn the jet supersonic. Ion velocities up to $2 \times 10^6\,\mathrm{m\,s^{-1}}$ can be assumed based on calculations. Indeed, direction-sensitive neutron detection can show that fusion reactions are occurring in and around the jet region [154]. In chamber 2 the incoming jet turns its kinetic energy into heat. The neutron imagery also shows that in the slowed and heated plasma accumulating in chamber 2 the large majority of neutrons are generated. The blue squares indicate windows for laser interferometry (LIF) measurements for examining the plasma state which have been conducted in cooperative experiment campaigns of LANL and VNIIEF.

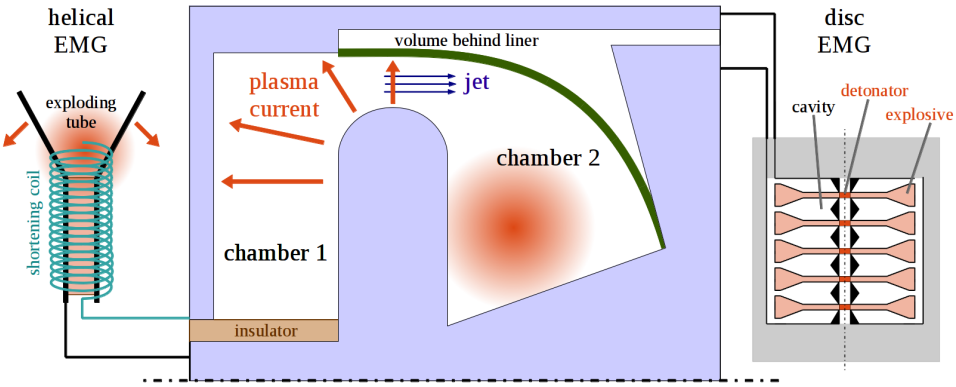


Figure E.13 Two-stage MAGO chamber with liner implosion
If it was possible to create magnetised plasma of a longer lifetime in the MAGO experiment, i. e. 5-10 µs instead of 2 µs, then it might be possible to achieve fusion burn ignition with a magnetically driven liner implosion in the second phase of the experiment. The accordingly modified experimental setup described by Garanin et al. [155] is depicted above. Here, the first EMG has also the purpose of feeding the current loop of the second one, a so-called disc EMG, sketched on the right, which can then deliver an energy of up to 200 MJ. The current flows around disc cavities which are imploded from the inside outwards compressing the toroidal B-fields in the cavities. The resulting current pulse peaking at ≈ 100 MA pushes the green liner forward with the force of the rising magnetic field in the volume behind it. The liner hits the central protrusion, closes chamber 2, and then compresses the magnetised plasma. In principle, if the compression was adiabatic and the gas only made of the intended D-T mixture, a compression ratio of 10 would suffice to bring the plasma of a MAGO chamber to the point of fusion burn ignition [270]. However, it seems this has not been realised experimentally yet, whereby the main problem is a too large energy drain and the too short lifetime of the magnetised plasma.

initially established magnetic field. It has toroidal field lines and as $|B| \propto \frac{1}{r}$, its magnitude decays with distance from the central axis of the rotationally symmetric geometry. This means that the restriction of the electron mobility in the drawing plane directions is least severe far away from the central rod. This helps to explain why the first plasma currents cross from electrode to electrode near the nozzle region and thus why most current paths encircle large parts of the gas volume in the left chamber where the magnetic pressure can build up.

In the course of the experiment the entire gas volume turns into plasma. The gas in the second chamber gets heated and ionised by the energy input from the impacting jet, while in the first chamber radiation plays a more important role for the spread of ionisation. The electromagnetic forces acting on the newly created plasma end up pushing almost the entire gas load of chamber 1 into chamber 2. Shock waves develop where the jet coming from the nozzle hits the stagnating gas. Due to the different properties of electrons and ions the shock waves feed plasma oscillations which can be picked up by the inductive coils. When strong electric fields exist in the shock waves and communicate the forces between the ion and electron systems, then part of these waves can be considered *collisionless*. Collisionless shocks delay the establishment of a local thermodynamic equilibrium and allow non-Maxwellian velocity distributions behind the shock.

Due to the highly dynamic and non-equilibrium conditions inside the MAGO chamber complex models have to be employed for accompanying simulations with the aim to match and explain the gathered diagnostic signals. Besides the neutron imagery these signals include the laser interferometry for measuring the plasma

density and the inductive probes sketched in figure E.12 as well as spectroscopy with diodes and filters ranging up to X-ray energies. After tuning model calculations and simulations for closely matching the measured data, Garanin et al. infer the following plasma conditions [155]: In the jet the ions reach velocities of 0.5-2 $\times 10^6\,\mathrm{m\,s^{-1}}$. This means deuterium ions have 40 and tritium ions 60 keV kinetic energy. This explains why fusion reactions can be induced when a fast ion hits a slow one from the outer region of the jet or one which has lost its energy by bouncing against the wall. The plasma created in chamber 2 has ion temperatures reaching $\approx$ 10 keV while the electrons stay colder with up to 2 keV (from the streaming gas in the jet where electrons and ions have similar velocities the lighter electrons take away less kinetic energy). The mean number density of the hot fusion plasma is around $10^{18}\,\mathrm{cm^{-3}}$.

The main problem of the experimental status reported in [155] is that the plasma becomes too cold too quickly. The reasons given besides normal heat conduction are instabilities arising from the interplay of thermodynamic and magnetic pressure and the large heat capacity of heavy atoms evaporating from the chamber walls. The plasma cools within 2 µs while the liner implosion with velocities reaching into the km/s range takes 10 µs.

General Fusion and the fast slow liner

Although the mainstream projects of fusion energy research are pursuing laser-driven ICF on the one and MCF[23] in tokamaks and stellerators on the other hand, it is often pointed out that MTF might represent a minimum in cost and technology demand situated in between. MTF involves pulsed implosions while at the same time relying on magnetic forces. As such it incorporates principles of both ICF and MCF.

Besides the solid liner techniques there are liquid liner approaches where magnetised plasma structures are compressed in imploding cavities surrounded by liquid metal. This has much more fundamental advantages than just ridding of the need of fabricating a steady supply of new liner cartridges to be smashed. If the fusion plasma is surrounded by liquid metal then all the neutrons and electromagnetic radiation emanating from the plasma get caught in it. The first main advantage is that the reactor structure will not be degraded by neutron irradiation. One other important point is that designs are possible where all the relevant portions of fusion energy output end up as heat transferred to the liquid metal. Liquid metals, with their high heat capacity and low viscosity, are ideal media for energy storage and transport. The energy can be easily carried out of the reactor core region and to a heat exchanger for powering a conventional water steam cycle. The last crucial advantage is that fusion fuel can be bred from lithium if it is part of the liquid metal mixture.

Because of these system-inherent advantages and because of the thought that with MTF no severe technical challenges remain unsolved and existing technologies just needed to be combined in the right way companies have been appearing conducting research using venture capital. An interesting example is General Fusion® founded by Michel Laberge in 2002 [163] who are working on realising the reactor

[23] *magnetic confinement fusion*, see below

concept explained in figure E.14. The steam piston drive and the acoustic shock implosion of the liner are the two key features of their reactor concept. Powering the pistons with steam means that the necessary energy can be taken in the form of heat directly out of the steam cycle without first converting it to electricity for feeding eventually wasteful machinery. This increases efficiency and lowers the threshold for break-even.

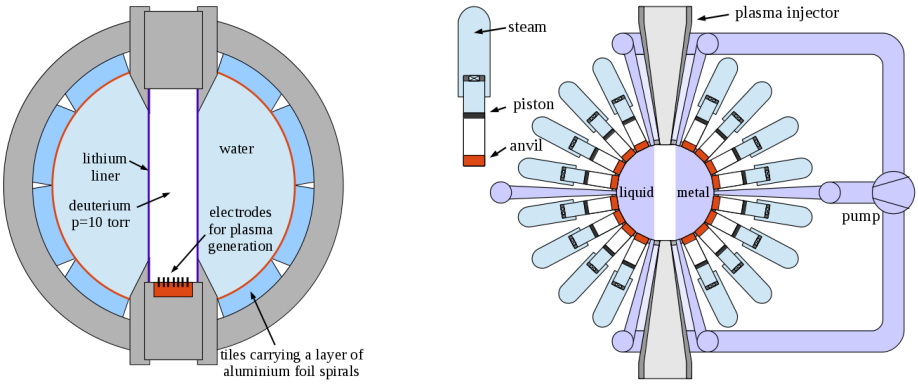


Figure E.14 General Fusion's liquid liner approach
The schematic on the left shows Michel Laberge's demonstration experiment [247]. The magnetised plasma target is a structure with U-shaped current lines created in deuterium gas at 10 torr near one end of the tube by a pattern of two concentric arrays of electrodes. It spreads upwards inside the tube because of the magnetic pressure tending to stretch and expand the current loops. The metal liner to be imploded around it is made of lithium, has an outer diameter of 29 mm and a wall thickness of 0.75 mm. That minimal thickness is needed so induced currents decay slowly enough. The liner tube is placed into a hollow sphere filled with water. Plastic foil separates the water from the lithium. The tiles on the inside wall carry spirals cut out from aluminium foil. When a 100 kJ current pulse from a capacitor bank is used to evaporate the aluminium foil patterns a concentric shock wave is created in the water. The lithium tube implodes with final velocities around and even above 4 km/s. The neutrons detected with scintillation counters outside the 8 cm thick metal wall of the sphere allow to infer a neutron yield of 2000 per shot. The sketch to the right shows the reactor layout envisaged by General Fusion. In the spherical reactor core volume liquid metal (Pb-Li) is kept streaming fast in a circle creating a vertex with almost vertical walls. The main point is that the implosions of the liquid metal liner are acoustically driven. This brings much higher implosion velocities within reach than piston-driven liquid liner implosions as in fig. E.15. Therefore, the piston-anvil assemblies surrounding the reactor core are a crucial part of the reactor design. They have the task of efficiently converting thermal energy from water steam first into kinetic energy carried by accelerated pistons and then into the kinetic and potential energy carried by an acoustic compression wave moving towards the centre where it will crush the plasma by imploding the cavity in the liquid metal. The plasma injectors are variations of *Marauder* type plasma canons developed in the early 90s within federal US research programs [110, 111, 210]. They create a pair of spheromaks which collide and merge in the very centre of the reactor core. The injectors have conical metal structures for compressing the spheromaks on their way. Ideally, the momenta of the merging spheromaks cancel out and the resulting spheromak is at rest with a lifetime of several microseconds giving a finite time window for staging the liner collapse.

With a demonstration experiment (also depicted fig. E.14) he and his team showed that a spherical shock wave in a spherical volume of liquid can implode a metal-lined cylindrical volume fast enough so as to achieve fusion in compression-heated magnetised plasma. An interesting technical challenge involved in this experiment is the manufacturing of quite thin-walled tubes of pure lithium which is as soft as modelling clay. But why choose lithium which on top chemically reacts violently with water and has to be carefully sealed during the test section preparation? The reason is that lithium has roughly half the density of water, thus the metal layer riding on the water shock front is not prone to Rayleigh-Taylor instability.

More recently, the General Fusion team have carried out experiments with a

sphere volume surrounded by 14 piston-anvil assemblies for tests of imploding a liquid metal vortex with a diameter of about 10 cm [387]. However, the free inner surface of the liquid lead does not stay smooth under the impact of the acoustic shock wave. Spikes, droplets, and spray are created and attributed to Richtmyer-Meshkov instability [447].

To outline some context for the MTF approach of General Fusion and to understand how their innovations are connected to previous research one can consider the following three technical aspects. As mentioned already in figure E.14 the conical plasma injectors which are able to create pre-compressed spheromaks go back to [110, 111, 210]. Previous research on spheromak merging can be found in [336]. Finally, the concept of imploding a vertex in rotating liquid metal has been explored in the 1970s [482, 484] at the US Naval Research Laboratory (NRL). The comparison with the NRL work (Turchi et al. [482, 484]) is particularly interesting. This work on directly piston-driven liquid liner implosions is explained in figure E.15. Turchi et al. showed experimentally that the inner surface of a liquid metal vertex being imploded can be stabilised through magnetic pressure. A reactor concept based on that technology was termed *LINUS reactor*.

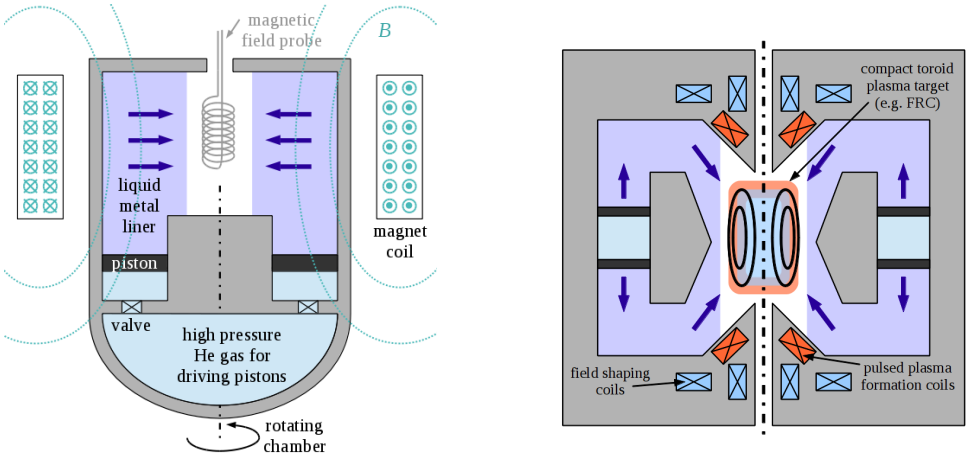


Figure E.15 The LINUS project at NRL
The US Naval Research Laboratory (NRL) had been conducting research on liquid liner implosions in the 1970s. The left sketch above shows the principle of a demonstration experiment which had proven that magnetic fields can stabilise the shape of the liner surface and keep it smooth. A reactor proposal is sketched on the right. The technical key features of the demonstration experiment are that the whole test section is rotating, that a freely moving annular piston pushes the liquid metal, and that fast valves make the connection to the high-pressure driver gas reservoir. But how does the magnetic field stabilise the free surface of the liquid metal? Again, on a fast enough time scale one can imagine the magnetic flux in the cylindrical cavity being conserved. This means that ripples on the surface protruding into the empty volume will be subject to an increased magnetic pressure and thus pushed back in line. It becomes also clear that the main structural material must be a dielectric material, since a conducting metal able to carry a substantial fraction of induced currents would distort these effects. In the reactor proposal depicted on the right the annular piston driving mechanism is made symmetric. A more detailed geometry suggesting how to make free annular pistons technically feasible is shown in [483].

Distinguishing the magnetic solid liner implosion approach from the piston-driven liquid liner concept, the first is sometimes referred to as *fast liner* whereas the latter is termed *slow liner* [522]. With that context the conceptual advancement step made by Laberge et al. becomes clearer: by introducing the anvil between pistons

and liquid and by transitioning to acoustic shock waves they made the slow liner concept substantially faster.

What are the technical challenges, besides the liner surface stability issue, lying still ahead with the General Fusion reactor concept? On the one hand the outlets above and below the vortex need to be closed during implosion to protect the plasma injectors. This task is complicated by the vacuum requirements. On the other hand the structure surrounding the reactor core volume is burdened at the same time with the mechanical loads from the acoustic waves and the corrosive effect of the hot liquid metal which makes it non-trivial to devise a system with long-term stability.

E.2.6 MCF - magnetic confinement fusion

Magnetic confinement means using magnetic forces for bending the trajectories of leaving particles back into the main plasma cloud. One challenge is that the magnetic field geometry has to accomplish this for both electrons and ions which differ by both charge and weight. Density and pressure in MCF reactors are relatively low because magnetic field strengths are limited for large scale devices. Therefore, the way towards energy break-even has to be made by lengthening the burning time. Constantly repeated plasma shots of several seconds duration or steady state operation need to be achieved in order to make MCF power plants possible. With extended time scales also those instability mechanisms with slower buildup characteristics have to be controlled. Consequently, another important challenge of MCF research is the topic of plasma instability.

The two common device classes for magnetic confinement are the *tokamak* and the *stellarator*. The first is explained in figure E.16. In the tokamak the particle trajectories are bound onto torus-shaped surfaces. These surfaces are layered on top of each other like the yearly wood layers of a tree. As the tokamak plasma carries a net current, some of its main instabilities arise from the pinch effect explained in figure E.17. The plasma current is an essential ingredient to the tokamak's magnetic field pattern, and in order to drive it a steadily rising transformer coil current is needed. Due to this transient coil current requirement tokamaks can only be operated in pulsed mode.

The stellarator concept goes without a plasma current contribution to the magnetic field which allows for continuous operation. Both toroidal and poloidal magnetic field components are supplied through external coils. The layers of closed surfaces in which particle trajectories are possible have principally the same torus topology as in a tokamak, but they look somewhat warped and deformed. Another great advantage of a stellarator with its current-less plasma is the freedom from current-driven instabilities.

MCF is the single mainstream research branch aiming at enabling the commercialisation of fusion power. The research reactors are multi-billion-dollar projects only feasible in the form of internationally coordinated and cooperative campaigns. In the tokamak line of research the *Joint European Torus* (*JET*, located near Abingdon, Oxfordshire, UK) is the largest machine that has been operated. The $200\,\mathrm{m}^3$ torus hosts a $90\,\mathrm{m}^3$ volume of DT plasma with the overall mass[24] of 0.1 g. The

[24]With $10^{-3}\,\mathrm{g/m}^3$ the density inside JET's plasma is a million times smaller than the density of

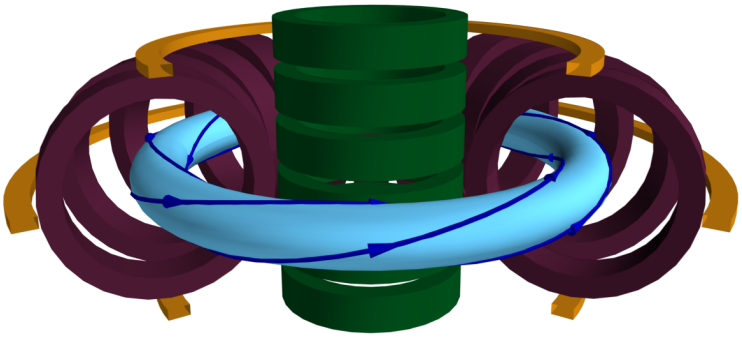


Figure E.16 Simplified tokamak geometry
In the tokamak setup the plasma is confined to a torus-shaped volume by strong magnetic fields. The plasma volume is depicted in light blue. Several types of coils are necessary for different purposes. The purple-coloured coils create a toroidal magnetic field. In principle, this forces the corkscrew trajectories of charged particles to circle around within the torus. However, there are drift forces counteracting the plasma enclosure, one of the reasons being the higher magnitude of $\vec{B}$ on the inside where the toroidal field coils are spaced narrower. The drift problem can be alleviated by introducing a net plasma current giving rise to a poloidal field component allowing the helical trajectories illustrated by the blue arrow lines. The centrally placed coils depicted in green serve that purpose. Steadily increasing their current will continuously feed the plasma current. Another secondary effect of the induced plasma current is heating. Tokamaks can only work in pulsed mode because the current through the central coils cannot be increased ad infinitum. Yet another set of coils, depicted in orange-brown, create a vertical field component and have the purpose to counteract the tendency of the plasma ring to grow in diameter, which is i. a. due to the plasma pressure. As large currents have to be maintained in the various coils, it is crucial to have them in the form of superconductors.

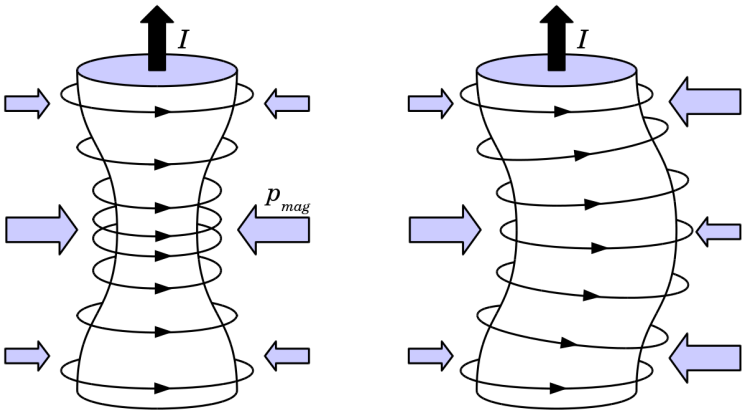


Figure E.17 The pinch effect
The current-bearing string of plasma in a tokamak is subject to the pinch effect. It pushes plasma away from regions of stronger and towards weaker magnetic fields. The example shown on the left is the sausage instability. Wherever the plasma tube narrows, due to the inverse proportionality of $|\vec{B}|$ with the radial distance r from the centre of the tube (containing the current paths), the magnetic field intensity rises. In the sketch this is symbolised by the narrower spacing of the field lines. The effect is a locally increased magnetic pressure around the neck. Thus the reaction of the increased magnetic field amplifies the initial perturbation. On the right, the case of the kink instability is shown. It arises from the narrower spacing of field lines near concave surfaces and the lowering of $|\vec{B}|$ above convex curvatures.

toroidal magnetic field is 3.45 T. Plasma temperatures up to 3×10^8 K (≈30 keV) can be reached. In 1997 a world record was set when transiently reaching 16 MW of fusion power (and the output-to-input power ratio $Q_\mathrm{p} = P_\mathrm{fusion}/P_\mathrm{heat} = 0.62$) as the peak in a shot of about 1 s duration. In another experiment, 4 MW were upheld over 4.5 s [222, 304, 351].

After JET, the next planned step in the tokamak branch is the *International Thermonuclear Experimental Reactor (ITER)* which is currently under construction in Cadarache, France. The aim is to demonstrate a steady and stable pulsed operation of igniting and burning fusion plasma with 500 MW of fusion power output with a power amplification factor $Q_\mathrm{p} = P_\mathrm{fusion}/P_\mathrm{heat} = 10$. As 80 % of the fusion power is taken away by neutrons and only the 20 % of energy carried by α-particles contributes to self-heating of the plasma, the demonstration of a self-sustainable heat balance ($Q_\alpha \geq 1$ with $Q_\alpha = P_\alpha/P_\mathrm{heat} = 0.2Q_\mathrm{p}$) is more relevant than the mere energy break-even $Q_\mathrm{p} > 1$. For ITER, $Q_\mathrm{p} > 10$ and $Q_\alpha > 2$ are envisioned for pulses of 400 s duration and $Q_\mathrm{p} > 5$ for longer pulses (aiming towards steady-state operation with advanced plasma control methods) [149, 351, 425]. The targeted boundary conditions are a toroidal magnetic field of 5.3 T, a pressure ratio β of 0.03 [122], and a mean plasma temperature of 2×10^8 K. Half a gram of plasma is to cover a volume of 840 m^3.

In the stellarator branch of MCF research the largest existing and the largest future reactors are *Wendelstein 7-AS* and *Wendelstein 7-X* (or simply W7-AS and W7-X). As in a tokamak, poloidally wound coils produce the toroidal $\vec{B}$-field. In its conventional form, the stellarator needs an extra set of external coils for supplying the poloidal field components in the form of conductors wound helically around the plasma tube. However, instead of adding an extra coil set, it is possible to create the very same field topology modifications by twisting each one of the primary set of coils in a certain way. This leads to the concept of a *modular coil set* (both geometry types are shown in fig. E.18). Wendelstein 7-AS was the first such *advanced stellarator*. It was operated from 1988 to 2002 by the Max Planck Institute for Plasma Physics in Garching near Munich and demonstrated long pulse (or "quasi-steady-state") operation with a mean electron density around $4 \times 10^{20}\,\mathrm{m}^{-3}$ at 2.5 T while the electron temperature was 0.35 keV [207]. The maximal temperatures achieved[25] were 1.7 keV for the ions and 6.8 keV for the electron system. The deduced maximal β-value was 3.4 %.

Previous stellarators of the conventional type had shown that small deviations of the coil and field geometry from the ideal shape have drastic consequences for the confinement properties [207]. In that context the transition to individually twisted modular coils and tailored 3D field geometries opens up a large design freedom that can be used to minimise particle drift and other confinement-antagonising mechanisms. Hence, geometry optimisation becomes a crucial part of the design of an advanced stellarator. Due to the limited computational power available in the 1980s, the W7-AS coil set was only "partially optimised", i. e. not all desirable optimisation goals were addressed. W7-X, in contrast, will be "fully optimised" [52].

air, which is 1.2 $\mathrm{kg/m}^3$ at 20 °C.

[25] during discharges at lowered densities by one order of magnitude less than the maximally achieved $n_e \approx 4 \times 10^{20}\,\mathrm{m}^{-3}$ [207]

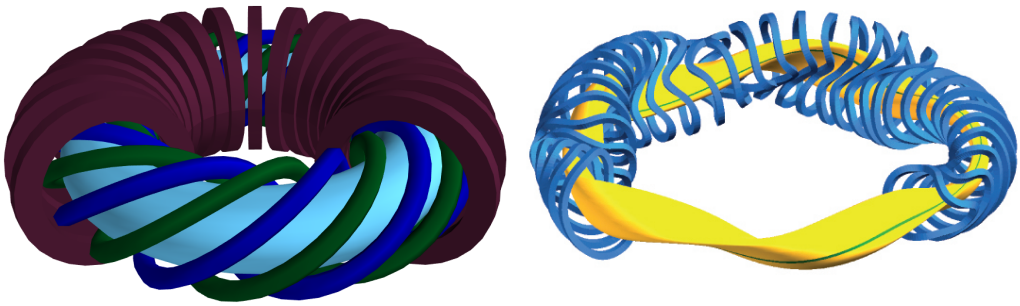


Figure E.18 The Stellarator
In a conventional stellarator the magentic field is created by two sets of coils as shown on the left. A primary set of coils generates the toroidal field. An additional set of helical coils (blue & green) furnishes poloidal field components. Two neighbouring helical coils carry currents in opposing directions. The great advantage is that a stellarator does not require a plasma current, so it can run in steady-state and is not prone to current-driven instabilities. With numerical simulations it is possible to find stellarator geometries where one single set of warped coils generates an equivalent plasma-confining field configuration. The picture on the right shows such a *modular coil set*. (Picture source for the modular type: the picture was created and released to the public domain by the Max-Planck Institute for Plasma Physics; publication via Wikimedia Commons).)

In the context of comparing the different confinement mechanisms, it should be noted that electric fields play a non-negligible role in modern stellarators. Hirsch et al. [207] note that improved stellarator confinement regimes could only be understood by updating the theoretical treatment as to incorporate the buildup of a radial charge imbalance due to orbit ions exhibiting a higher loss rate than electrons. Consequently, ions are kept on their helical torus tracks not only due to magnetic forces, at the same time they rotate around a line of accumulated negative charge.

E.2.7 Where sonofusion fits in

Sonofusion would be fusion occurring in the plasma sparking up inside imploding cavitation bubbles. A centimetre-sized cluster of micrometre-sized bubbles collapses, and in each imploding bubble a supersonic shock compresses and heats a small fraction of its vapour and gas content. Due to the smaller size, SF can be seen as one further step, as an extrapolation of the way from the H-bomb to LCF. The amount of available fusion fuel is unproportionally low, as the compression begins with gas, not solid or liquid hydrogen. On top comes the dilution of the fusion fuels among other atom species being present in the bubbles' vapour content. The scientific interest in SF so far is not the ignition of self-propagating fusion burn or the maximisation of an energy gain or payoff, but lies merely in the question whether some detectable amount of fusion reactions can be triggered at all. The odd thing about SF is the low energy density of the driving mechanism. While the H-bomb is driven by a fission bomb, and LCF by mankind's biggest laser machine, SF is induced by the low energy density of sound in liquid. This raises the question about the type of energy concentration mechanism thought to accomplish the task: it is the physics of the sequence of spherically imploding bubble clusters, collapsing bubbles, and concentric supersonic shocks inside the bubbles which is outlined in chapter 1.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description (of the quantity)
A	nucleon number
a	radiation constant; Wigner-Seitz radius; escape velocity (of particles leaving a plasma)
a_A	coefficient for asymmetry term of the Bethe-Weizsäcker formula
a_C	coefficient for Coulomb term of the Bethe-Weizsäcker formula
a_S	coefficient for surface area term of the Bethe-Weizsäcker formula
a_V	coefficient for volume term of the Bethe-Weizsäcker formula
$\vec{B}, B$	magnetic field
C_p, C_V	heat capacity at constant pressure/volume
c	speed of sound/light
c_p, c_V	heat capacity per volume at constant pressure/volume
E	energy
E_B	binding energy
$\vec{E}, E$	electric field
e	elementary charge
F	force; energy multiplication factor (plasma enery balance)
$h, \hbar$	Planck constant ("quantum of action")
k	Boltzmann constant
L	luminosity
M, m	mass
N	neutron number; particle number
N_{Av}	Avogadro number
N_D	plasma parameter (Debye number)
n, n_e, n_i	particle/electron/ion density
P	power
p	pressure
p_e, p_i, p_r	pressure of electron/ion/photon population
q	eletric charge
Q	reaction enthalpy
Q_α, Q_p	power ratios
$\mathcal{R}$	reaction rate
r	radius
T	temperature
T_e, T_i, T_r	temperature of the electron/ion/photon population
U	internal energy, potential energy
u	internal energy per volume, e. g. plasma energy density
u_e, u_i, u_r	internal energy per volume of electron/ion/photon population
V	electric potential; volume
$\vec{v}, v$	velocity
$\vec{x}, x$	position vector, spatial coordiante
Z	proton number
z	axial coordinate

List of Greek quantity symbols

Symbol	Description (of the quantity)
β	pressure ratio
Γ	coupling parameter
γ	adiabatic index
δ, δ_0	energy of the pairing term of the Bethe-Weizsäcker formula
ϵ_0	vacuum permittivity (electric constant)
ε	energy conversion factor in plasma enery balance computations
λ_B	de Broglie wavelength
λ_D	Debye length
$\lambda_{De}, \lambda_{Di}$	species-specific Debye lengths for electrons/ions
μ_0	vacuum permeability (magnetic constant)
ϱ	density
σ	(reaction) cross section
σ	Stefan constant
τ	time period/scale
τ_d	disassembly time
τ_{KH}	Kelvin-Helmholtz time scale
ω_p	plasma frequency
ω_{pe}	electron plasma frequency

List of particle symbols

Symbol	Particle
D	deuteron $= p + n$
e^-, e^+	electron, positron
p	proton
n	neutron
ν	neutrino
T	triton $= p + 2n$

List of abbreviations

Abbreviation	Description
EMG	explosive magnetic generator
FRC	field-reversed configuration
gnd	ground
IC,ICF	inertial confinement (fusion)
IEC, IECF	inertial-electrostatic confinement (fusion)
ITER	International Thermonuclear Experimental Reactor
JET	Joint European Torus
LANL	Los Alamos National Laboratory
LC,LCF	laser confinement (fusion)
LIF	laser interferometer

LINUS	slowly-imploding liner (as campaign name probably picking from the terms "slowly-imploding liner" and "liquid-metal cylindrical annulus" [301])
LLNL	Lawrence Livermore National Laboratory
MAGO	magnetic compression (магнитное обжатие, МАГО)
MC,MCF	magnetic confinement (fusion)
MTF	magnetised target fusion
NIF	National Ignition Facility
NRL	US Naval Research Laboratory
PFX	Penning Fusion eXperiment
POPS	periodically oscillating plasma sphere
RF	radio frequency
SEMF	semi-empirical mass formula
SF	sonofusion
SL	sonoluminescence
VNIIEF	All-Russian Scientific Research Institute of Experimental Physics (Всероссийский научно-исследовательский институт экспериментальной физики, ВНИИЭФ)
W7-AS	Wendelstein 7-AS
W7-X	Wendelstein 7-X

Appendix F

Nuclear binding energies – physical background

On a sunny day you can look up into the sky and witness a very safe and reliable nuclear fusion reactor. But don't look directly into it, its radiation is very strong and can damage your eyes. Luckily, one big advantage of the setup of our world is that the outer layers of the sun, the distance between the sun and us, the magnetic field of the earth, and finally the incredibly precious and thin layer of our atmosphere protect us against the more unhealthy parts of the big fusion reactor's radiation.

What drives this reactor? It is differences in nuclear binding energies. Knowing about the binding energies of atomic nuclei helps to understand nuclear reactions. Nuclear binding energies are determined by the laws of quantum physics, and in that framework they can be computed with a high precision. Yet, a simple and in many cases efficient formula for calculating the binding energy of a nucleus with a total number of A nucleons of which Z are protons is the *Bethe-Weizsäcker formula*[1], also called the *semi-empirical mass formula*:

$$E_B = a_V A - a_S A^{\frac{2}{3}} - a_C \frac{Z^2}{A^{\frac{1}{3}}} - a_A \frac{(A-2Z)^2}{A} - \delta(A,Z) \tag{F.1}$$

with

$$\delta(A,Z) = \begin{cases} -\delta_0 \cdot A^{-\frac{1}{2}} & \text{if } Z \text{ and } N \text{ both even} \\ 0 & \text{if } A \text{ odd} \\ +\delta_0 \cdot A^{-\frac{1}{2}} & \text{if } Z \text{ and } N \text{ both odd} \end{cases}$$

This formula is based on a simple droplet model and enriched by some end results of a quantum-mechanical description of the atomic nucleus. The five terms of the Bethe-Weizsäcker have this physical background:

Volume term: It is the first ingredient of the droplet model and reflects the presence of each nucleon in its neighbour's potential well. The attractive force between nucleons being responsible for the potential wells is the *strong nuclear force*, and one of its properties is that it is of very short range (only circa

[1] According to Pfalzner [347] the formula was first outlined by Carl-Friedrich von Weizsäcker [502], and later improved by H. Bethe, R. Bacher, and E. Wigner.

10^{-15} m [347]). Because of the short range, not every nucleon is in every other nucleon's potential well, each nucleon interacts only with a subgroup among all the other nucleons within the nucleus. This is the reason why the term is proportional to A and thus the volume, instead of being proportional to A^2. This is not much different from a droplet of water where also only a limited number of other molecules fit into the neighbourhood of one given molecule.

Surface term: This is the second ingredient of the droplet model, and it describes the surface tension. The thought is that a particle's transition from the surface to the inside of the droplet leads to an energy gain because it enlarges the number of neighbour interactions. Hence the tendency to minimise the surface area which is responsible for pulling free droplets together into a spherical shape, be it water, oil in water, quicksilver, or nuclear matter. Surface tension is also a precondition for a droplet being able to be excited into shape vibration modes. The connection between volume V, droplet radius R, and nucleon number A being $V \propto A \propto R^3$, the proportionality with respect to the droplet surface S transforms according to $S \propto R^2 \propto V^{\frac{2}{3}} \propto A^{\frac{2}{3}}$.

Coulomb term: It reflects the fact that protons with their positive charges repel each other. Neutrons, by contrast, don't interact with the Coulomb field. The electrostatic potential is inversely proportional to distance, which explains the proportionality with $A^{-\frac{1}{3}} \propto R^{-1}$.

Asymmetry term: The Schrödinger equation yields the wave equations for quantum-mechanical particles. Each stationary eigenfunction of the Schrödinger equation is also associated with an eigenvalue, the energy level of the state, and if particles are confined to a finite-size potential well then there will be finite distances between the energy levels. On top of that, protons and neutrons are fermions, and unlike bosons fermions are not allowed to cover the same wave function. This rule is called the Pauli principle, and it means that each energy level can be covered by a maximum of two identical particles which still have to differ in spin in order not to cover one and the same wave function. By consequence, the more nucleons there are in the core, the more energy levels will have to be populated. The only thing is that protons and neutrons are two sorts of particles and as such each type has an own independent pool of wave functions and energy levels. They are sharing the same potential well offering (besides an offset due to the Coulomb potential) practically the same energy levels for both types, but these energy levels are populated by the two sorts as if in two parallel dimensions. That means with 20 nucleons of one single type the energy levels would have to be filled from the bottom up through the 10^{th} level, while if there are 10 protons and 10 neutrons, only 5 levels will have to be filled on each side. Therefore, symmetry between protons and neutrons maximises the nuclear binding energy.

Pairing term: Since each wave function and energy level can be inhabitated by two fermions, the addition of every second fermion costs a bit more energy than the steps in between (when filling the entire nucleus' potential well one by one in thought experiment). This jumping between two energy cost levels

is added to the smooth function expressed by the other four terms, and in combining the two fermion types the above expression is yielded.

The coefficients of the semi-empirical mass formula need to be derived by fitting the model to a database of known[2] binding energy values. Chowdhury & Basu [86] used the Ame2003 database of binding energies by Audi et al. [16] and deduced[3] the following set of values in MeV: $a_V = 15.409$, $a_S = 16.873$, $a_C = 0.695$, $a_S = 22.435$, $\delta_0 = 11.155$.

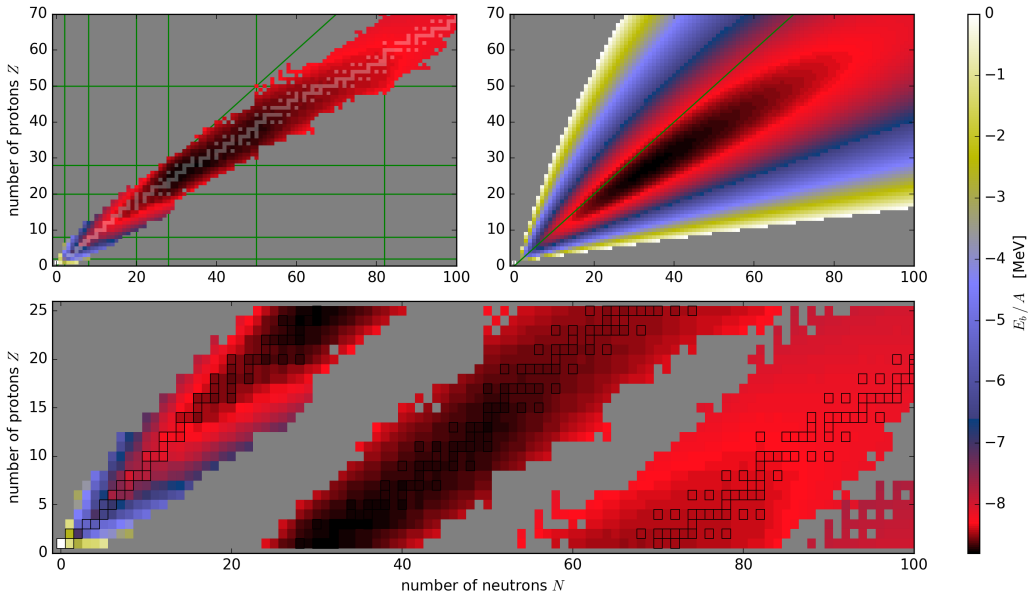


Figure F.1 Nuclear binding energy as function of Z and N.
The nuclear binding energy is depicted over Z and N in the form of a colour map over a nuclide table. The top row of diagrams shows a comparison between the empirical data of the Ame2003 database [16, 500] on the left and the semi-empirical mass formula (evaluated with the coefficients indicated in the text) on the right. The diagonal line traces symmetric nuclei with $N = Z$. It can be seen that the Bethe-Weizsäcker formula gives a good prediction of the general shape of the valley of stability with its tilt to the right towards $N > Z$ for heavier nuclei. Stable isotopes are indicated through the white shading in the upper left map. It can be seen that particularly long horizontal and vertical sequences of stable isotopes coincide with the magic numbers indicated by the underlying green lines. The lower diagram gives a closer view undisturbed by shading. For the sake of visibility the lower diagram involves a periodic vertical wrapping of the field and all three views are limited to $Z \leq 70$ and $N \leq 100$.

The Bethe-Weizsäcker formula gives approximate values for nuclear binding energies. It not only enables to tell whether a nuclear reaction will be exothermic or endothermic, but also indicates the general shape of the binding energy as a function of Z and N. For the lightest nuclei the formula however fails. Figures E.1 and F.1

[2]While molecular binding energies have to be deduced from careful caloric characterisation of reactions, nuclear binding energies offer a conceptionally easier way to get to know them because the energies invested or lost are so high that they translate into clearly measurable weight differences according to $E = mc^2$. It is called *mass defect* when an atom is by $\Delta m = E_b/c^2$ lighter than its separate ingredients.

[3]Chowdhury & Basu [86] present in fact four alternative sets of coefficients. The distinctions stem on the one hand from including or neglecting extrapolated entries within the Ame2003 database and on the other hand on the usage of two different objective function formulations for the least squares minimisation.

illustrate the general match while also pointing out the differences in detail. The figures compare the formula results with the Ame2003 database [16, 500] of empirical values. In reality, the binding energy as a function of Z and N is less evenly distributed as the Bethe-Weizsäcker formula would predict. There are additional features like the peaks and concave bends visible in the bottom right of figure E.1. There are special proton and neutron numbers leading to elevated binding energies and stable isotopes where the simple droplet model would not let expect them. In a 2D nuclide map such as figure F.1 this leads to longer than average sequences of stable isotopes in a row or column. Only a quantum-mechanical description of the wave functions for protons and neutrons in terms of a shell model involving spin-orbit coupling can explain some of these features. The *magic numbers* of elevated binding energies correspond to completely filled shells above which there is a larger energy gap to the next level. The magic numbers are 2, 8, 20, 28, 50, 82, and 126. They are depicted in the upper left of figure F.1 as horizontal and vertical green lines.

Radioactivity = downhill jumps into the energy valley

Talking about radioactivity in general, the 2D plot of figure F.1 can be very helpful. Radioactivity always stems from exothermic reactions, where the decay of an instable atomic state releases stored energy. In the case of γ radiation the isotope does not change, it originates only from a nucleus transitioning from an excited state (e. g. proton and neutron droplets vibrating against each other) into an energetically lower state sending off a high-energy photon. But for other forms of radioactivity the isotopes change which corresponds to jumps on the 2D isotope map, in particular downhill jumps into regions of lower potential energy.

In the case of β decay, β-radiation in the form of electrons or their antimatter counterparts, positrons, is released. But what happens inside the core? During a β^--decay one neutron of the core gets transformed into one proton, one electron, and an electron-type antineutrino. Hence, this corresponds to one horizontal step to the left and a vertical step upwards, and it can only occur to isotopes on the lower side of the energy valley. In the other version, the β^+-decay, a proton changes into a neutron, a positron, and an electron-type neutrino. This diagonal step to the lower right occurs to isotopes on the upper side of the valley.

Another example is the emission of α-particles consisting of two protons and two neutrons. It corresponds to two diagonal steps towards the lower left. Logically, this happens most often to the heaviest nuclides on the far end of the field and tends to occur a bit more on the upper side of the tilted valley.

Nuclear binding energy differences are driving stars – and earth

In today's world view the expanding universe stems from an initial big bang. It is assumed that during its first moments, the young universe consisted of highly concentrated energy which then condensed into matter (and antimatter), and that this primary cooling process yielded as the base matter for the birth of the first stars a cloud of hydrogen and helium. The question then is where all the heavier elements came from which we find on earth. The solution to the riddle is that our solar

system must belong to a secondary generation of stars. In the late part of their life cycles stars transition from hydrogen burning to the burning of helium, then carbon and further heavier elements in a sequence of collapses connected with temperature increases. The assumption is that super novae blow substantial parts of that heavy matter into the open space, and that these blast waves can even trigger the compression and gravitational collapse of neighbouring primordial hydrogen clouds. Our solar system is imaginable as the consequence of such a sequence of events.

The logical next question is about the origin of elements heavier than iron and nickel, which we also find on earth. If stellar fusion burn describes the steps from hydrogen over helium, carbon etc. towards the most stable isotopes, why should the process go upwards again on the energy scale on the far side of the nuclide map? The answer is that the death of a star in the form of a super nova corresponds to a fast cooling and quenching of the fusion burn reactions, at least for the star's blown-away outer layers. Neutron capture can lead to heavier nuclei while the plasma becomes too cold for a consuming burn.

What does this mean for the energy household of nature and mankind? It first means that all burnable fuels are finite, stars burn until all dense hydrogen clouds will be consumed. However, the time scale of billions of years of this process is not a relevant time frame for human decision-making. Restricting our view to life on earth it still means that all energy sources can be traced back to the initial fuel inventory furnished by the big bang. If we build nuclear fusion reactors, we burn light nuclei fuel as stars do. If we run nuclear fission reactors, we burn down stocks of stored energy frozen into heavy atoms when the burn of a preceding generation of stars was interrupted. If we use geothermal energy we tap into the heat stream generated in the inside of our planet from radioactive decays of heavy nuclei, i. e. this taps into the same stored energy stock. The chemical burn of oil, gas, and coal is an analogue consumption of temporarily stored *stellar* energy, or more precisely *solar* energy from *our* sun. Considering *renewable* energy sources at last: if it is carbon-based it can be described as short-time (in-cycle) biomass-based storage of solar energy, wind and water streams stem from solar energy stored (during very short time frames) in mechanical way as potential and kinetic energy, and only photovoltaic or solar-thermal power plants tap directly into the incoming solar energy stream.

It means that anything running in this universe runs because of the initial kick supplied by the big bang. Planets are orbiting because everything hasn't collapsed yet into a single final black hole. All energy consumption has to be traced back to burning nuclides down towards $^{62}_{28}$Ni. The gradient of nuclear binding energies is the only available energy source in this universe.

Respecting the precise physical meaning of the words, one has to say that there are no energy sources nor dissipators. Energy is conserved. Energy can only be transformed from one form into another. In such a wording one has to say: things in this universe can run as long as entropy can be easily maximised, and this means until all highly concentrated spots of energy have been well distributed throughout the universe in the form of low-temperature thermal radiation. The question is just: will this slow cool death scenario be pre-empted by a gravitational re-collapse of the universe? And can one ask what happens "after" the re-collapse? Or beyond the big bang?

Getting back to the earth-bound discussion of energy sources and consumption, we can state that all energy source options relevant to mankind at this point have two things in common: they all trace back to the same origin of nuclear binding energy gradients, and they all have severe disadvantages, at least if we want to exploit them at an amount necessary if billions of people are to live a 20th century northern hemisphere lifestyle. What the energy options do not have in common are the types and combinations of disadvantages: radiation, radioactivity, land use, impacts on atmosphere and climate, finiteness of stock among others. And as we are talking about human group and swarm decisions, it is perhaps not too negligible to note, that scientific input can and should be given not only with a narrow technologic view in mind, but also considering what history and curious analysis can yield about the properties and actions of human societies. Otherwise there will always be Babylonian confusion and what one person means with terms like "infeasible" or "dangerous" is not or wrongly understood by too many other persons.

Perhaps one can nevertheless point out one distinction which is the CO_2 non-neutrality of burning fossil biomass in comparison to all other energy sources, whereby playing with the climate also means that fossil fuel burning is the only option risking global catastrophe and not only local catastrophic impacts.

List of symbols

Symbol	**Description**
A	nucleon number
a_A	coefficient for asymmetry term of the Bethe-Weizsäcker formula
a_C	coefficient for Coulomb term of the Bethe-Weizsäcker formula
a_S	coefficient for surface area term of the Bethe-Weizsäcker formula
a_V	coefficient for volume term of the Bethe-Weizsäcker formula
c	speed of sound/light
δ, δ_0	energy of the pairing term of the Bethe-Weizsäcker formula
E_B	binding energy
m	mass
N	neutron number
R	radius
S	surface
V	volume
Z	proton number

Appendix G

Neutron sources

G.1 Samples of radioactive materials

There are two possibilities: neutrons as a byproduct of spontaneous fission and neutrons emitted by a nucleus after having received an impacting α particle.

Fission: Heavy isotopes like californium ($^{252}_{98}$Cf), which are unstable and go into energetically lower states through spontaneous fission, can be used as neutron sources if they have a long enough half-life so they keep on decaying over a decent period of time. The products of a nuclear fission do not only comprise the two lighter halves of the initial nucleus, but also one or more neutrons and gammas, and they all carry away variable amounts of energy. The fission products are often radioactive as well, emitting photons to rid themselves of excessive energy and to go into the ground state, and emitting electrons and antineutrinos in a β^--decay to release excessive neutrons. The resulting additional hard γ spectrum is a disadvantage of that type of neutron source.

α radiation-induced neutron emission: It is possible to create instable excited nuclei by pumping them up with an impacting α particle. The lighter the target nucleus the lower is the coulomb barrier for the impacting α particle to overcome. Elements like lithium, beryllium, or aluminium can be used as target. They form a compound nucleus together with the α, which often enough decays by giving off a neutron. Thus, neutron sources can be manufactured by combining α emitters like $^{239}_{94}$Pu, $^{241}_{95}$Am or $^{242}_{96}$Cm in a small vessel with the target materials. In the case of the often used PuBe neutron source, Plutonium-239 decays spontaneously with a half-life of 24 110 years sending off α particles with an energy of 5.15 MeV, and the $^{9}_{4}$Be together with the α forms[1] a $^{13}_{6}$C* (which would be a stable isotope in the ground state), which decays emitting one neutron with up to 11.0 MeV. Such neutron sources also emit gammas, but at much lower energies compared to spontaneous fission materials, which is their main advantage for lab usage [344].

[1] Mostly a ^{13}C atom is formed, but not always. There are other channels involving different reactions contributing with smaller probabilities to the neutron spectrum of the PuBe source, see [491].

G.2 Pulsed neutron generator (PNG) based on D-T fusion reactions

Nuclear fusion reactions are a source of neutrons. The difficulty with fusion lies in the Coulomb barrier repelling the reactants. If fusion reactions are to take place in a heated plasma, this translates to the high temperatures $> 10^8$ K, that make fusion so difficult to harness in a laboratory setup. However, giving deuterium (and tritium) ions enough collision speed for triggering D-D (or D-T) reactions is no big deal for modern particle accelerator technology. Small and portable sealed accelerator tubes have become commercially available, where D or T ions are accelerated to hit a target containing D or T as well, which can produce up to $10^{11}\frac{n}{s}$ of 14.1 or 2.45 MeV in continuous or pulsed mode.

G.3 Generation of photoneutrons by a high-energy electron beam

Photoneutrons are neutrons emitted from a nucleus which has been put into an excitated state by a γ photon. The necessary γ rays can be furnished by bremsstrahlung from electrons. This is the short story of how a beam of fast electrons impacting on a target can be used to generate neutrons emanating from that target. A little more precise image of the phenomenon can be given in the framework of quantum mechanics and with the help of figure G.1.

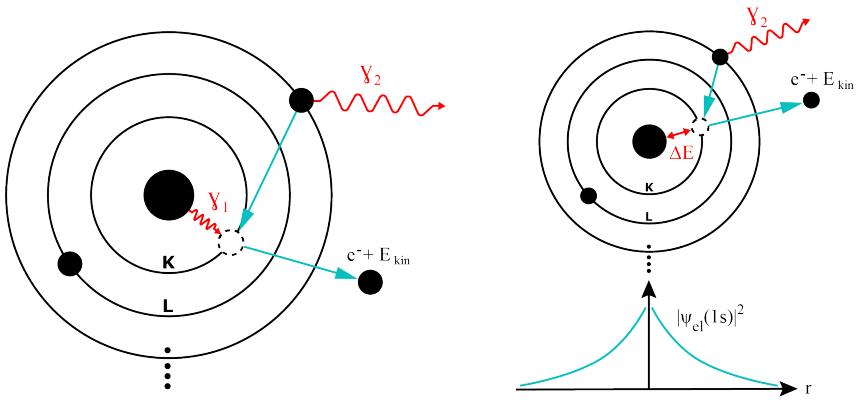


Figure G.1 Generation conversion electrons
Sometimes an excited atomic nucleus emits an electron and an X-ray photon instead of a γ photon. In the thought approach shown in the left image the initial γ photon always leaves the nucleus, and then, as an option on the way leaving the atom, it is absorbed by an electron close to the nucleus where the electric field is still strong (a free electron without the help of an outside electromagnetic field cannot absorb a photon, there can only be Compton scattering), and this kicks the electron out of its orbital. Finally, the free orbital in one of the inner shells is filled with a valence electron from a shell further outside, and that transition entails the emission of the X-ray photon γ_2 with much lower energy. This can explain the two experimental observations of sharp characteristic X-ray bands and the generation of positive ions. In the alternative image depicted on the right hand side, the γ never leaves the nucleus, it is instead a purely virtual photon serving as a quantum-mechanical exchange particle transmitting energy from the nucleus to an electron whose wave function overlaps with the nucleus so it has a nonzero probability density inside it. The latter image is more powerful since it explains why only these X-ray bands are observed which correspond to s-orbitals of the innermost shells being filled, why there are no X-rays from $l, p, ...$-orbitals which have zero probability density in the atom's centre.

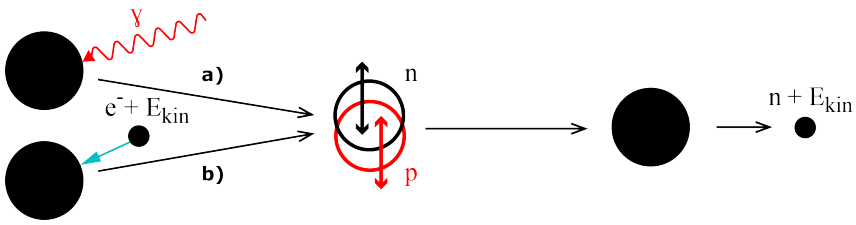


Figure G.2 Generation of photoneutrons
A nucleus can absorb a photon (left, a) because via dipole interaction it can be excited into a state of dipole oscillation where the system of protons swings against the ensemble of neutrons (middle). Neutron emission is one channel of de-excitation from that excited state (right). When a fast electron ($> 7\,\mathrm{MeV}$) traverses an atomic core, the excited state can also be reached (left, b), and in the formalism of quantum mechanics the transition in the moment when the wave functions of electron and nucleus overlap can be explained again with the help of a virtual photon as carrier of dipole interaction.

The term "photoeffect" or "photo-electric effect" originally stems from describing the ability of light shining on a metal surface to supply the energy to pull electrons out of their place in the conducting valence bands within the metal and let them fly away into vacuum where they can be collected and observed in a detector. One observes that not the light intensity (correlated with the electric field) is decisive for pulling electrons out of the metal block's potential well, but that rather the wavelength is the decisive threshold. That contradiction with classical physics makes it a historically important experiment. Only the introduction of the equation $E = h\nu$, where h is Planck's constant[2] lifts the contradiction, and this is a consequential step because it states that the energy carried by electromagnetic waves comes in quanta of h proportional to the wave's frequency, which finally leads us to assume that light waves must have also a particle-like nature. "Photoeffect" now generally describes interactions of light in its photon nature when it lets other quantum-mechanical particles forming matter transition between states with different energy: photons can dump their energy into e.g. chlorophyll molecules or solar cells by lifting electrons into excited states, or inversely, photons leave a neon tube when electrons drop back into their ground state. Consequently, the field between a radio emitter and an antenna, and also between two magnets, can be seen as made up by innumerable photons. The concept of energy and force transmission by real or virtual carrier particles is very common in modern physics.

Photoneutrons are emitted from a nucleus that had been lifted into an exited state after the reception of a photon (a γ photon with an energy above $\approx 7\,\mathrm{MeV}$). The photons may have come from far away, or, alternatively they may have been mere virtual carrier particles of the electromagnetic field, and this latter option is what allows the whole kinetic energy of an electron to be used in one single step for the excitation of an atomic nucleus, and it may lead to the emission of a neutron with up to the same kinetic energy from that atom. Figure G.2 outlines the process. The neutron flux emanating from the target is isotropic, as any angular information is forgotten by the excited nuclei during their short existence. The neutron energy spectrum is influenced by neutrons scattering with other nuclei before leaving the target material.

[2]It used to be called "quantum of action", invoking the notion that the effect of light comes in quanta of h and avoiding yet the particle interpretation of photons.

Lists of symbols and abbreviations

List of quantity symbols

Symbol	Description
a	radiation constant
E	energy
h	Planck constant (“quantum of action”)
$K, L, M, ...$	electron shell labels
ν	frequency, e. g. of electromagnetic radiation
$s, p, d, ...$	orbital labels
Ψ	quantum-mechanical wave function

List of particle symbols

Symbol	Particle
α	alpha particle consisting of $2p + 2n$
D	deuteron consisting of $p + n$
e^-	electron
γ	photon
p	proton
n	neutron
T	triton consisting of $p + 2n$

Appendix H

Bubble nucleation

A steaming pot of boiling water and the fog cloud forming above it or a drink with slowly melting ice cubes or slowly growing ones in the freezer are examples of continuous phase transitions in thermodynamic equilibrium. Both phases are present but separated by a boundary, and particles transition across the boundary from one state to the other. Quick and unsteady phase transitions far from equilibrium can also occur if there is a metastable state to begin with and one phase is not present at first. Some examples:

- clouds forming in clear air initially oversaturated with water vapour after the introduction of dust particles serving a nucleation sites for condensed droplets (a way for volcanoes and aeroplanes to manipulate weather and climate),
- condensation in a cloud chamber where flight paths of charged particles are the sites of droplet nucleation,
- a bottle of cleanly filtrated beer cooled below freezing point can freeze through in a minute after the first ice crystal has been nucleated in the subcooled liquid, and small concussions can suffice for that,
- water in a clean cup can be brought into a superheated state in the microwave, and it begins to boil instantaneously when e. g. a teabag touches the water surface and furnishes the impurity necessary for bubble nucleation.

In particular the cloud chamber is very interesting: this tool, invented about a hundred years ago, to visualise particle tracks stemming from radioactivity or cosmic rays exists also in the inverse direction, as bubble chamber, invented 1952 by Donald A. Glaser [166, 167]. In a bubble chamber[1] a liquid is brought into a superheated state by quick isobaric heating or by pressure lowering through pistons, and subsequently the first incident charged particles can draw their tracks as chains of nucleated boiling bubbles until everything boils up bringing the ensemble back to thermodynamic equilibrium and making the chamber unsensitive again.

[1] The bubble chamber represented a great progress in particle physics because particle interactions are proportional to density and the liquid in a bubble chamber is much more dense than the gas in a cloud chamber. Glaser received the 1960 Nobel prize in physics.

The principle of bubble nucleation by particle scattering can also be used for triggering acoustic cavitation. Why it leads to bubbles better suitable for SF experiments than other ways of nucleation, will be described below.

literature: [165–167, 176, 217]

H.1 Tension creates superheated liquids

A sound wave of amplitude 0.1 atm in air leads to the absolute pressure oscillating between 0.9 and 1.1 atm. A sound wave of 10 atm amplitude lets the pressure range from −9 to 11 atm. This second case can however only exist in liquids and solids, where particles are under the attractive influence of their neighbours. In a gas there can be no negative pressure just as there can be no negative density (and particle momenta[2] cannot show into backwards directions). In a solid, a state of tension is nothing special, and it is e. g. part of the sound wave travelling through a block of steel after it has been hit by a hammer. In a liquid, it is however something more exotic because of a lack of purity under normal conditions. The example here is seawater where the centrifugal forces of a vortex (e. g. behind the wings of a ship propeller) or the implosion wave following the expansion wave of an explosion lead to water foaming up under cavitation. The preexisting bubble nucleation sites consist of solid particles and microbubbles of noncondensible gasses. Even if these are not present and one has water in a homogeneous liquid phase, a high content of dissolved gasses can also prevent the occurrence of states of large tension, think of soda sparkling upon lowering the pressure. But in pure liquids with very low dissolved gas content states of tension can be created. Any state with the pressure being below the saturation vapour pressure is a metastable superheated state, and the vapour pressure above any liquid is positive. Tension ("negative pressure") corresponds to strong overheating.

H.2 Bubbles and bubble clusters equilibrating tension

In principle the tension can be pushed to the limit where the liquid ruptures, i. e. where the tension overcomes the attractive neighbour forces responsible for keeping the liquid phase together. It can be instructive to follow the rupturing liquid in a thought experiment. At the moment when the tension overcomes the attractive forces, the collective phenomenon of surface tension can be imagined having no influence. Can this lead to unregularly shaped cracks propagating through the liquid? Not really, because as soon as a cavity exists, the pressure inside it is zero at first, and it rises to the saturation vapour pressure soon while being filled with particles evaporating from the surrounding liquid into the gas phase. This keeps the surface tension in action, and instead of cracks it is manifold structures of bubble clusters that spread throughout the liquid [298]. These structures can resemble organic tissues with their branching, fractal-like properties, or Lichtenberg figures. They are the result of the time-variation of tension in the liquid volume, which is a result of the interplay of elasticity, pressure equilibration through bubble growth, and inertia

[2]Momentum transport from many reflected particles is what creates gas pressure on a surface.

limiting the communication speed of pressure signals. In the collapse of bubble clusters the same forces are at play, implosion fronts propagate into the clusters, and these effects are able to generate very different boundary conditions for the collapse of different bubbles in different regions of the cluster [326].

H.3 Critical radius, bubble growth instability

A liquid is an unordered, temporally changing structure of particles where distances among neighbours vary and can be described by distribution functions. Describing empty space between particles, beyond which size should it be described as a bubble, and what determines whether a bubble will grow or instead shrink and disappear again as so many other *empty space fluctuations*? Assuming e. g. the potential well of the Lennard-Jones potential, it implies that when neighbours are pulled out of each others' potential wells the inter-particle forces first become stronger but then diminish quickly. The other picture is the collective macroscopic consequence of the same thing, the surface tension. That implies that when the forces across the bubble space have decayed to negligibility, the surface tension is the right model to describe the force that acts to close the cavity again. Viewing a spherical bubble with radius R as two half spheres, it can be thought that the two halves are pulled together by the surface tension σ acting along the ring where they touch. The resulting force is σ multiplied by the length of this line, i. e. $F_{\text{ring}} = \sigma L_{\text{ring}} = 2\pi R\sigma$. For a bubble in equilibrium where the internal pressure p stands against the force aiming to close the bubble, the internal pressure can be obtained by relating that force with the size of the surface circumscribed by the ring line:

$$p = \frac{F_{\text{ring}}}{A_{\text{ring}}} = \frac{2\pi R\sigma}{\pi R^2} = \frac{2\sigma}{R}. \tag{H.1}$$

Here, the external pressure of the surrounding liquid has been left out of the consideration. The (offset) pressure p inside bubbles in equilibrium is inversely proportional to the radius. For a round microcavity filled with nothing but vapour, where the internal pressure is constant and given by the vapour pressure, this means that if its radius is smaller that a critical radius

$$R_c = \frac{2\sigma}{\Delta p}. \tag{H.2}$$

the surface tension will win and manage to collapse the bubble again and let it vanish while the vapour content gets condensed. In equation H.2 the external pressure has been taken into account, so $\Delta p = p_i - p_{ext}$ is the pressure offset. But if the liquid is supersaturated and the cavity has a radius larger than R_c, then the surface tension loses against the vapour pressure being larger than the external pressure, and the bubble grows. It has to be kept in mind, that both, Δp and σ are functions of the temperature T.

H.4 Means of bubble nucleation

In a liquid with a distribution of preexisting impurities (like microbubbles of non-condensible gasses or gas-filled crevices in dust particles) the pressure can never go

far below the vapour pressure. Every lowering of the pressure can be annihilated by bubble growth. This is undesired if one wants to have the biggest cavitation bubbles for a given sound pressure. These are achieved if the liquid is put under tension and the cavitation bubbles snap open at once. Under such conditions inertia will lead to an overshooting of the bubble size. The two common options are laser cavitation and scattering of charged particles. Both techniques allow a precise control of the nucleation time. In both cases the radiation source is external, the radiation is transported into the region of interest without material interaction. Laser cavitation also allows the precise control of the region of interaction: the energy deposition takes place where the light beam is focused and the $\vec{E}$-field becomes sufficiently strong for ionisation. With good optic equipment the size of the ionised region in the direction across the beam can be lowered to match the order of the wavelength used, i. e. hundreds of nanometres for visible light and microns for infrared light. In the longitudinal direction the resolution is larger and depends on the focus angle.

The other method is the trace of ionised matter left behind by a fast-travelling charged particle. Observing the traces of cosmic rays in a bubble chamber, the tracks can go from one end of the chamber to the other. This is different with neutrons where fast ions are a secondary effect stemming from neutrons undergoing elastic scattering with an atomic nucleus. The advantage of relying on neutron scattering for cavitation nucleation is that one has a small region of interaction (the short path of the knocked-on ion) and that, as in the case of laser-induced cavitation, a noninteracting transport mechanism (the neutron) is bringing the energy into that region. The only compromise is that the exact location of neutron scattering is not controllable, it will happen at random locations throughout the volume covered by the neutron beam and lead to cavitation wherever the liquid tension is strong enough and the energy density deposited by the knock-on ion high enough. Whereas laser interaction ionises the liquid within an elliptical volume, fast ions lead to ionisation along a narrow track. The bubbles are created in chains along the track (see e. g. the photographies in [166]). The energy loss rate the ion experiences along its track, $\mathrm{d}E/\,\mathrm{d}x$, is dependent on the ion mass, charge, and speed, and on the material it traverses. The track lengths for small ions like deuterons or α particles vary from microns if the initial kinetic energy is a fraction of a megaelectronvolt up to millimetres for kinetic energies of several MeV.

A simplified estimation approach consists in saying that a bubble is formed whenever the necessary energy W_c for building a bubble with a radius larger than the critical radius is being deposited within a piece shorter than $L = 2R_c$ along the trajectory of a fast ion. This however misses the fact that the conversion of the fast ion's kinetic energy to immediate localised heat in the liquid is not always near 100 %.

H.5 The energy cost of bubble nucleation

It has been said that in a metastable superheated liquid any bubble grows as soon as its radius is larger than the critical radius $R_c = 2\sigma/\Delta p$ with $\Delta p = p_i - p_{ext}$. According to Gibbs [164], the cost of forming a bubble of critical size ("by a reversible

process") is the sum of surface and displacement work

$$\begin{aligned} W_c &= E_{\text{surf}} + E_{\text{vol}}, \\ W_c &= A\sigma + V\Delta p, \\ W_c &= 4\pi R_c^2\sigma - \frac{4}{3}\pi R_c^3(p_i - p_{ext}). \end{aligned} \tag{H.3}$$

The second term is negative since we are considering the case $p_i > p_{ext}$. If the liquid is under tension, there is already an energy gain alone from displacing the surrounding liquid. Using relation H.2 equation H.3 becomes

$$W_c = \frac{16\pi\sigma^3}{3(p_i - p_{ext})^2}. \tag{H.4}$$

If one wants to be more exact, more terms can be considered. In [33] the energetic cost is given as

$$\begin{aligned} W_c &= E_{\text{surf}} + E_{\text{vol}} + E_{\text{vap}} + E_{\text{kin}} + E_{\text{fric}}, \\ W_c &= 4\pi R_c^2\sigma - \frac{4}{3}\pi R_c^3(p_i - p_{ext}) + \frac{4}{3}\pi R_c^3\varrho_m H_v + 2\pi\varrho_l R_c^3\dot{R}^2 + E_{\text{fric}}. \end{aligned} \tag{H.5}$$

where H_v is the molar evaporation enthalpy and $n/V = \varrho_m$ is the molar density. The added terms are the vaporisation enthalpy of the bubble content, the kinetic energy of the surrounding liquid at the moment the critical radius is surpassed, and E_{fric} is the work lost on viscous friction in the liquid. The paper [33] points at the same time to references where proofs of the negligibility of the last two terms in the case of water and organic liquids can be found. Different authors use different formulae [33], e. g. Seitz [411] and Ing et al. [217] use

$$W_c = E_{\text{surf}} + E_{\text{vap}}. \tag{H.6}$$

H.6 Energy conversion and empirical nucleation parameters

When a fast ion loses energy on its trajectory through matter through Coulomb interactions, then most of the energy is transferred put into excited states in the system of electrons. The energy can end up in the form of e. g. dissociated molecules, fluorescent light, or just plain heat. The *thermal spike* theory of bubble nucleation [411] says that bubbles are formed by that fraction of a fast ion's lost energy which is transformed into heat quickly and locally. As that fraction is material-dependent and because the microscopic processes determining it can be quite complex, in practice one often speaks of abstract conversion factors which are deduced from empirical data. This means transitioning from the simplified threshold assumption that the critical energy must be supplied within a track length corresponding to the critical diameter

$$W_c = 2R_c\frac{\mathrm{d}E}{\mathrm{d}x} \tag{H.7}$$

to the similar equations

$$W_c = \eta 2R_c \frac{\mathrm{d}E}{\mathrm{d}x} \tag{H.8}$$

$$W_c = kR_c \frac{\mathrm{d}E}{\mathrm{d}x} \tag{H.9}$$

$$W_c = L \frac{\mathrm{d}E}{\mathrm{d}x} \tag{H.10}$$

which are the defining equations for the introduction of parameters like a *nucleation efficiency* η [10], a *nucleation parameter* k [217, 413], or simply a threshold interaction length L [33, 217].

List of symbols

Symbol	**Description**
A	surface area
E	energy
$\vec{E}, E$	electric field
η	nucleation efficiency
F	force
H	enthalpy
k	nucleation parameter
L	length
p	pressure
R	bubble radius
ϱ	density
σ	surface tension
V	volume
W	work, energy
x	distance, spatial coordinate

Appendix I

Research on sonofusion and cavitation resonators by Lahey, Saglime, Cancelos et al. at RPI

Soon after the publication by Taleyarkhan et al. in 2002 [458] a team at RPI including R. T. Lahey Jr., R. C. Block, Y. Danon, F. Saglime III, and S. Cancelos started attempts to repeat the ORNL sonofusion experiment independently at the Gaerttner Laboratory, the linear accelerator (Linac) lab of RPI. The goal was to use the pulsed neutrons created at the accelerator's target to cavitate deuterated acetone and check for the two signatures of eventual D-D fusion reactions, 2.45 MeV neutrons and tritium production. They started out with resonators assembled mostly from glass parts furnished by the group of Taleyarkhan at Purdue University and continued with their modifications of that resonator type as well as new design variations. Unfortunately, the best resonator was the old one from ORNL and thus they did not report a successful detection of sonofusion signatures [250, 390].

First, the principles and results of the conducted sonofusion experiments will be described below. Thereafter follows a review of important technical aspects of the works done at RPI. This includes the pulsed neutron source, the resonator manufacturing process, the resonator features, and the development tracks of the resonator designs.

I.1 Sonofusion trials at the Gaerttner Laboratory

The setup

For the sonofusion trials of Frank Saglime et al., published in [250, 252, 390], resonators № 1 and 4 of figure I.5 were used and installed in a freezer which could be lifted into one of the neutron beamlines of the Gaerttner Laboratory at 25 m distance from the neutron source. While during characterisation the resonators were filled with normal acetone (C_3H_6O), for SF trials this was replaced by deuterated acetone (C_3D_6O). The resonators were characterised by (a) recording amplitude and phase frequency sweeps of the sound pressure and (b) vertical sound pressure

profiles such as the example shown in figure I.1. The hydrophone described in appendix O.1 had been used. For (a) the hydrophone had been fixed in the centre of the liquid volume at a vertical position identified before as the desired pressure antinode. The profiles (b) were made through consecutive small vertical shifts of the hydrophone position while adjusting the frequency for tracking the desired resonance while it shifted slightly due to the perturbation by the hydrophone (see page 41 in [390]). The hydrophone measurements had to occur with the top head removed. With the top head and piston in place, no sound pressure measurement was possible any more, but the consistency of hitting the right resonance was ensured by keeping the top piston at an equal height as the free liquid surface and could be checked by the right spatial distribution of cavitation bursts. For setup and characterisation cavitation bursts could be induced by using neutron radiation from a plutonium-beryllium (PuBe) source. A LabView® code was written for the purpose of continuously adjusting the frequency to maintain the oscillation mode at resonance when the resonance frequency shifted slightly over time. The control was based on the driving voltage phase lag between the amplifier input and the transducer, and by analysing this phase lag only at times of smooth and periodic vibration when undisturbed by cavitation bursts. This automatic resonance maintenance control was an important improvement in comparison to the experiments by Taleyarkhan et al. at ORNL where it was done by hand. During long hours of run the tracked resonance frequency covered an interval of 150 Hz.

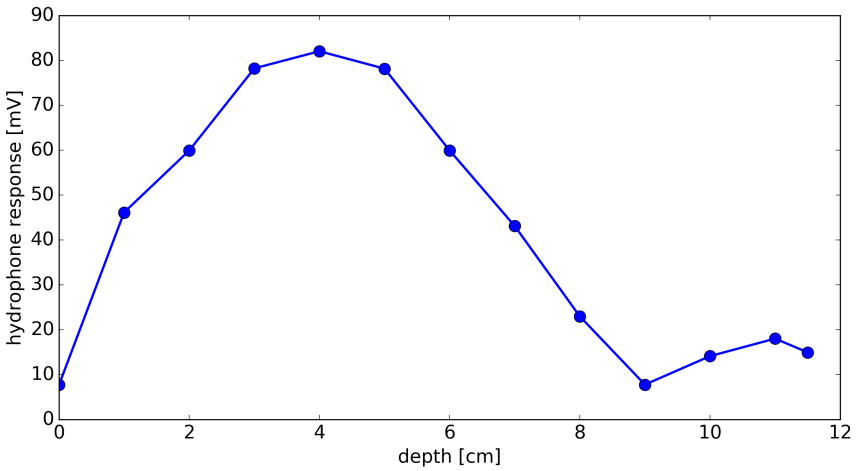


Figure I.1 Pressure profile of resonator № 1.
This plot represents the data of figure 21 in [390] (p. 41). The position of the hydrophone tip, the depth, is counted from the free surface of the liquid. The frequency was 19.9 kHz.

With that setup sonofusion trials were conducted in the Linac's pulsed neutron beamline with a bare tantalum target bombarded by 50 MeV electrons as the neutron source. In many other details the experiment protocol of Taleyarkhan was followed (see appendix D), in particular with respect to the temperature setting (between −5 and 5 °C) and the choice of the driving voltage amplitude. Under the assumption that neutron-induced cavitation can occur only if the acetone is under a tension stronger than −7 bar and that the electronics and the resonator are working in the

linear range, it was aimed at operating at a sound pressure amplitude of $\sim 15\,\mathrm{bar}$ ($\sim 30\,\mathrm{bar}$ peak-to-peak) by doubling the driving voltage after having reached the cavitation threshold.

The instrumentation for neutron detection consisted of an organic liquid scintillator, a photomultiplier, and the electronics for neutron-gamma pulse shape discrimination. Tritium decays under emission of beta radiation, that is electrons. The short range of electrons in matter makes the detection of this radioactivity somewhat more difficult. The measurements of the tritium content were therefore done by direct mixing of samples of acetone with scintillation cocktail. For these measurements the tritium samples were sent to an independent laboratory (New York State Wadsworth Laboratories) that specialises in radiation measurements.

Looking for neutrons while cavitating with the PuBe neutron source

A few SF trials had been conducted with the PuBe neutron source positioned near the resonator for cavitation nucleation. PuBe produces a feature-rich neutron spectrum ranging up to 10.7 MeV [344] plus α and γ radiation. The fact that the neutrons emanate isotropically and at random times from the radioactive sample means that cavitation is induced whenever a neutron scatters and deposits enough energy in the acetone where and whenever the tension is sufficiently large. This means the bubble size distribution is suboptimal in comparison to pulsed neutrons time-matched with moments of largest tension (see appendix H). With the goal of minimising the influence of other time-varying influences from the lab environment, neutron spectra were recorded during several repetitive cycles of the acoustic drive turned on for one minute and then turned off for the same amount of time. With acoustic drive on cavitation rates between 18 and 27 bursts per second were observed. The two-minute cycle was repeated ten times. By that time, a total of 10^6-10^7 events, mainly coming from the PuBe source, had been recorded, and according to the author's estimate this ensured that a conservatively estimated lower limit amount of produced fusion neutrons of $10^3\,\mathrm{s}^{-1}$ would not be missed by the detection system. Thus, two neutron spectra were obtained by summation. The comparison of the cumulative spectrum over all "on" times with the one representing all "off" times yielded however no statistically significant differences in the region around 2.45 MeV. The spectra[1] are congruent without noticeable differences.

Looking for neutrons while cavitating with the pulsed neutron beamline

The pulsed neutrons allow matching possible times of neutron scattering and bubble cluster nucleation with moments of highest tension in the liquid. The resonators were placed in the freezer, kept at $\sim 0\,^\circ\mathrm{C}$, and positioned in the neutron beamline. The beam cross section is several inches wide. The scintillator for neutron detection was positioned off to the side, but due to a lacking fusion neutron signature in the statistics it was moved closer and ended up being touched by the beam during the documented experiments. Gamma (γ) radiation from the neutron source would saturate the detection system and thus a two-inch layer of lead was put in the beam

[1] plot on page 89 in [390]

abolishing that problem. The neutron pulse from the beam did not lead to a system dead time and posed no problem due to the separation in time from the moment of bubble collapse. Two versions of control experiments were possible in this setup: switching on or off the acoustic drive while the chamber was in the pulsing beam and tuning the neutron pulse in and out of phase with the tension in the liquid. Both levers switch cavitation on and off. The "in-phase" setting of the neutron pulse was determined empirically as the phase angle leading to the highest cavitation rate (20-30 bursts per second).

The neutron results were shown in the form of two histograms of neutron counts over time, where the time axis was the time since the beam trigger. The two histograms[2] compared the cases of acoustic drive "on" versus "off" and showed cumulative data taken over 300 seconds. No difference is visible between them.

Looking for tritium after long cavitation with the PuBe neutron source

The second tracer of fusion is the generation of tritium of which tiny concentrations can be detected due to its radioactivity producing β radiation. For the purpose of producing a measurable buildup of tritium in the detuerated acetone, long-term cavitation runs were conducted. The PuBe neutron source had been employed to avoid the high costs of using the accelerator over many hours of time. Cavitation runs of 12 hours with resonator Nº 1 (the chamber assembled from parts sent from Purdue) and 24 hours duration with Nº 4 (RPI chamber) were documented. The RPI team sent 16 samples of 1 ml of deuterated acetone to an external laboratory (NYS Wadsworth Lab) which had state-of-the-art equipment making them capable of detecting tritium at concentrations as low as 17.3 pC/ml. The samples were labelled with random numbers and thus examined in a random order. One quarter of samples was directly out of the bottle of fresh deuterated acetone. One quarter was put in the resonator and subjected to the radiation of the PuBe radiation source for 24 ours but without cavitation. One quarter of samples came from the 12 hour cavitation run with resonator Nº 1, and the last one from the 24 hours run with resonator 4. The plotted results[3] correspond to the accumulated counts over the four samples in each case during 300 minutes of measurement time for each sample. Significantly, the count rates seem to be substantially elevated (several $\sqrt{N}$ distance) for the cavitated samples in comparison to the two control sample batches. But this elevation has been described as statistically nonsignificant due to uncertainties related to the sample masses originating from unskilled pipette usage. Thereupon the same experiment had been repeated and all sample quantities taken by pipette were measured on a precision balance which allowed a proper normalisation of the radioactivity measurements with respect to the sample mass. This second set of tritium data, which had come in too late for being included for publication in [250, 252, 390] yielded negative results [393].

[2] figure 66 on page 91 in [390]

[3] figure 67 on page 92 in [390] or figure 37 on page 51 in [250]

Observed temperature limit of cavitation

Lowering the temperature of the liquid lowers the degree of superheating at a given degree of tension and at the same time increases the surface tension. As a consequence the energy threshold for bubble nucleation, W_c, increases. While good cavitation rates were still seen at $-5\,°C$, cavitation was not possible any more at $-20\,°C$ [393].

I.2 The nature of neutrons used for cavitation induction

As described in appendix H, particle scattering processes are an ideal tool to allow cavitation bubble nucleation because neutron pulses impacting at well-defined times are a perfect way for creating clusters of bubbles in liquids under large tension whereby the bubbles can grow to a large size quickly while staying rather empty. While the SF trials of Taleyarkhan [458] were relying on 14.1 MeV D-T fusion neutrons from a PNG, the RPI team worked with a different external neutron source: photoneutron pulses produced by the impact of an electron beam on a tantalum target. What differences does that make?

The Gaerttner Lab hosts a travelling wave linear accelerator driven by 9 klystrons producing a pulsed beam of electrons with energies up to 60 MeV and pulse widths between 5 and 5000 ns. Aiming this electron beam at a target containing heavy atoms, the high-energy collisions of electrons with atomic nuclei can be used to generate a flux of photoneutrons emanating isotropically from the target. The two substeps involved in the process are (1) electrons colliding with nuclei and emitting bremsstrahlung and (2) the photons engaging in photo-nuclear reactions and inducing the production of photoneutrons (this step is explained in more detail in appendix G.3). As there are small delay times involved in the primary reactions, the neutron production decays with the gamma flash from the electron impact, so the neutron pulse shape looks similar in time to the shape of the electron pulse (at least sufficiently close to the target, where time-of-flight differences of neutrons with differing energies have not played out yet). The accelerator target room is shielded with concrete and an earth mound, except for three small outlets. These holes in the shielding are forming what can be used as neutron beamlines traversing the neighbouring buildings hosting the experiment setups. Large sections of the neutron paths consist of evacuated steel tubes which minimise the beam interaction with air.

The energy spectrum of the neutron beam can be described as a sum of several weighted contributions of the form $\Phi(E) = E/E_T^2 \exp(-E/E_T)$, where $E_T = kT$ is a temperature parameter. A parameter fit formula in good agreement with empirically gained flux data has been devised in [391]. That spectrum is plotted in figure I.2. It has the bulk of neutron energies between 0.05 and 0.5 MeV and becomes thin beyond 2 MeV. In principle, neutron energies up to the maximum electron kinetic energy are possible, but the probability of a neutron having a relatively high energy is low. In fact, 90 % of the neutrons are below 2.5 MeV and 99 % are below 9 MeV.

An RPI report for NYSERDA[4] [68] documents simulations that had been dedicated to estimate the rate at which neutrons from different sources manage to

[4]New York State Energy Research and Development Authority

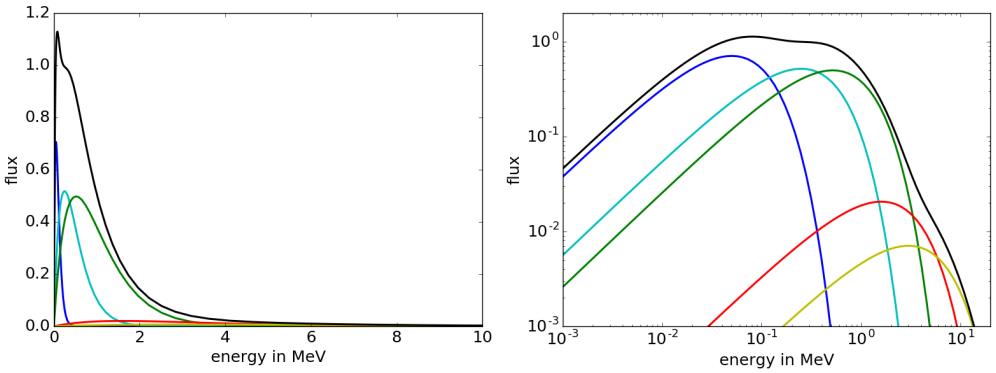


Figure I.2 The neutron energy spectrum of the RPI Linac.
The neutron spectrum generated by the linear accelerator at the Gaerttner Laboratory when a beam of 60 MeV electrons hits a bare tantalum target is an evaporation spectrum, and it can be described by a superposition of five contributions of the form $\Phi(E) = \sum_i w_i \Phi_i(E) = \sum_i w_i E/E_{T_i}^2 \exp(-E/E_{T_i})$ [391]. The coloured lines show the five contributions with characteristic thermal energies $E_{T_i} = 0.05, 0.25, 0.52, 1.60, 3.00$ and with weights $w_i = 0.0960, 0.03520, 0.7039, 0.0896, 0.0576$. The black line is the sum, and it mirrors the probability density function (PDF) of neutron energies.

nucleate bubble clusters in the SF resonators. MCNP [307] simulation results were postprocessed to yield the statistics of neutrons colliding with other nuclei in the region of interest at the centre of the resonator geometry. SRIM [538] simulations gave the stopping powers $\mathrm{d}E/\mathrm{d}x$ in deuterated acetone of the three possible nuclei that can be knocked out of a C_3D_6O molecule. With the statistics of knock-on nuclei and their initial energies the statistics of bubble nucleation could be gained. For an assumed tension of -15 bar in the region of interest, the computations were made for the three types of neutron sources: 14.1 MeV, 2.45 MeV, and the RPI Linac neutron spectrum. The findings were that the bubble nucleation was dominated by knock-on deuterium nuclei, the contributions from C and O were relatively small in the case of all three neutron spectra, and that the deuterons created most of their bubbles near the end of their flight after having been slowed down to less than 1 MeV. Together with the fact that a decent bubble burst rate could be observed experimentally [390] with the resonators placed in the Linac's neutron beam, this implies that the neutron source type should not be of much relevance. However, before drawing too many conclusions, some potentially problematic aspects of this study should be considered. Firstly, the assumption made for the Linac's neutron spectrum was very different from [391], secondly, the bubble nucleation model was somehow simplistic because a bubble nucleation efficiency of 100 % (i. e. $\eta = 1$ in equation H.8) has been taken. Lastly, for the critical energy W_c the defining equation H.6 neglecting the volume work had been chosen without further discussion.

Figure I.3 shows the ion stopping powers computed for this report with SRIM. It can be seen that deuterons distribute their energy more thinly along their track than the heavier ions. On the other hand the lighter deuterons can receive much more energy from a neutron when hit centrally than the heavier nuclei.[5] This has the consequence that, given a low enough critical bubble nucleation energy, deuterons

[5] Upon central hit a neutron tranfers 44 % of its kinetic energy onto a deuteron, whereas for carbon and oxygen nuclei the ratio is 2.4 % and 1.4 %.

can create the much longer bubble chains.

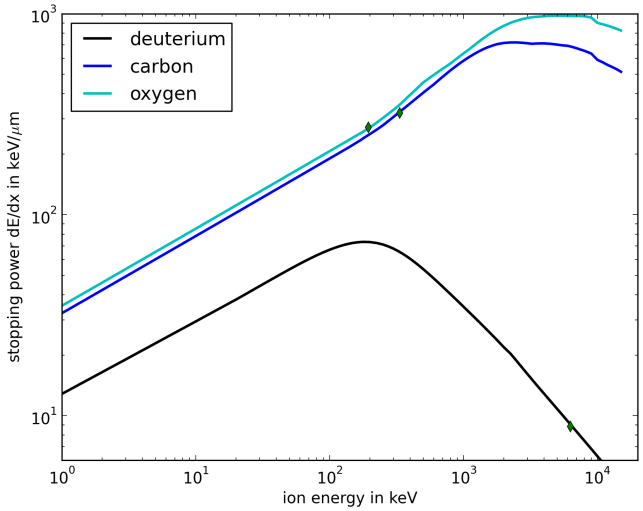


Figure I.3 The stopping powers of fast D, C, O ions in deuterated acetone.
These lines show the energy loss per unit of trajectory segment length of the three types of possible knock-on nuclei, deuterium, carbon, and oxygen. The green diamonds indicate the maximal energy that can be received after a central hit from a neutron with 14.1 MeV. Starting out with these energies the deuteron will still be able to propagate over about 0.4 mm distance, while the heavy ions will only be able to cover $\sim 2\,\mu$m.

So, in order to be in the position to draw some telling conclusions about what relevance the type of neutron source has on the SF experiment, one would need to know and explain what the initially chain-shaped bubble clusters turn into over several acoustic cycles.

I.3 Resonator manufacturing process

The acoustic resonator of the type introduced by [505] and sketched in figure I.4 is in principle a liquid-filled glass flask. The glass parts are made through glassblowing. The first assembly step consists in attaching the piezoelectric transducer to the main glass cylinder. After electric leads have been soldered[6] to its electrodes, the piezo ring is glued to the glass main cylinder with two-component epoxy resin. Different kinds of epoxies have been used for the older resonators, whereas in the later RPI models only Stycast®-1264 was used and cured at room temperature [69]. The choice had fallen on Stycast-1264 for its low viscosity and relatively long curing time. It had been tried to minimise the width of the gap between glass and transducer, and as this width had come down to 0.5 mm, the most suitable way of fixing the transducer,

[6]Suitable soldering techniques are described by the transducer manufacturers. Channel Industries® recommend the following [353]: (a) cover the lead with solder beforehand, the recommended solder is Sn-62 (62 % tin, 36 % lead, 2 % silver), (b) cover the soldered lead with noncorrosive flux, (c) melt solder on the tip (a small one) of the soldering iron of approximately 30 Watts, (d) press the lead down on the electrode with the soldering iron until the solder flows and keep the soldering time short to prevent dissolving the electrode.

elaborated by S. Cancelos [69], has been found to be this protocol [70]:

- The soldering joint on the inner electrode surface of the piezo ring has to be made very well and with as little elevation as possible.
- Begin with a band of adhesive tape going horizontally around the glass cylinder and continue with rolling on the adhesive tape, so that many layers pile up. The edge of each next layer has to match the one below, so that the side of the pile forms a surface and this surface can become the setting jig for the piezo ring.
- The gap between the transducer and the glass should have the same thickness all around the cylinder. Therefore, the rotation axes of glass flask and transducer have to be aligned well. Three straight pieces of wire can be useful here if stuck into the gap.
- In order to prevent the leaking of the low-viscosity epoxy at the lower rim of the piezo ring, more tape is added to close the slit between the piezo ring and its jig.
- The epoxy resin is filled into the gap with the help of a syringe with needle. The resin's viscosity has to be low enough so it passes the syringe needle and gets spread out at the bottom of the gap due to gravity.
- The gap has to be filled up to the top rim carefully, so that there will be no air bubbles left. The curing process of the epoxy has to be slow enough for allowing the necessary handling time.
- The alignment wires have to be pulled out early enough while the epoxy is still not cured. In case the jig-fixation is not completely tension-free, they have to be left in as long as possible, so the resin's raised viscosity can help conserve the transducer's alignment.

After the transducer has been cemented into place, the lower piston can be glued in using silicone, thus sealing the flask's bottom outlet at the same time. The comparison of resonators examined at RPI (see figure I.5) shows that more flexible as well as more tight setups, even involving cast metal forms, have been tried for that connection. For the fixation of the piston counterpart in the top half, also other techniques avoiding silicone glue connections have been tried out. On resonator no. 3 it has been a clamped rubber hose and on resonator no. 5 clamping by a rubber O-ring squeezed by a teflon screw.

Finally, the resonator flask has to be closed and sealed by connecting the top head to the main glass cylinder. This connection can be made through a simple silicone bead, which is the method given in Taleyarkhan's sketch (fig. D.1 or [461]). The samples in figure I.5 show that this can be done with more or less excess silicone on the outside covering larger or smaller surface areas on the glass wall. What cannot be seen in the pictures and what is not really specified in Taleyarkhan's sketch is how far the top head is pushed down into the fresh silicone, i. e. what gap width

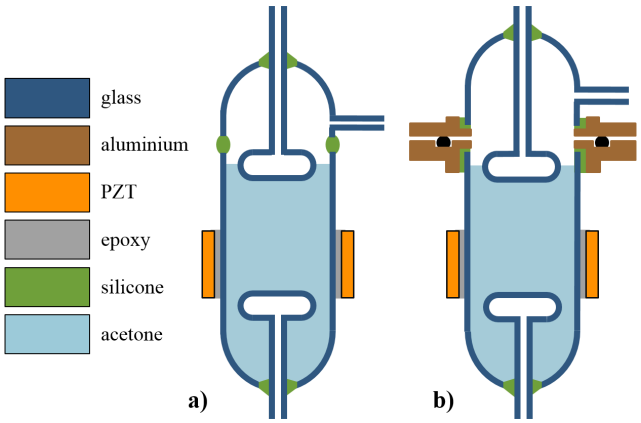


Figure I.4 Resonator schematic.
This diagram shows schematics of the two versions of acoustic resonators that have been in use at RPI for the sonofusion experiments reported in [250, 390]. The basic design goes back to the work of West & Howlett [505]. In the version on the left the top head is sealed to the main cylinder through a silicone bead, this design corresponds to the one described by Taleyarkhan et al.. In the version on the right the two glass parts are connected to aluminium flanges with silicone. The flange parts are held together by bolts. This allows much easier and quicker opening and closing of the chamber. The sealing occurs through a rubber O-ring squeezed between the aluminium surfaces. As the connection between the glass and the flange parts is made of a thin silicone layer covering an area along the outside of the cylinder, and as the thin layer of silicone is backed by a metal part, the acoustic coupling between the resonator parts is much stronger in that case compared to the other version.

remains between the two glass rims. (It will be seen that this affects the acoustic coupling between the two parts.)

To put the resonator to use, it has to be connected to the lab infrastructure in three ways, by electric cables for the voltage supply of the transducer, mechanically to hold it in place, and by hoses for the purposes of pressure control and degassing. Whereas sufficiently thin electric cables can be deemed to contribute only negligibly to the acoustic coupling of the resonator structure with the environment, this is not the case for the hosing, e. g. if heavy material is used as seen on resonator № 3 in figure I.5, and also not for the resonator support construction, in particular if parts of the resonator are clamped (picture of resonator 3) or the aluminium flange is even bolted onto a holding device (resonator 4, bolted onto a metal fork, as also visible in figure 19 (b) on page 39 of [390]).

I.4 History of resonators examined at RPI

Within the bubble fusion research project at RPI, several resonators had been in use, either for characterisation and preliminary experiments looking only at cavitation rates, or for sonofusion trials.

Description of the single resonators

Figure I.5 shows the resonators having been used at RPI, which will be described in the following. They had all been already assembled at RPI before the collaboration with Karlsruhe (KIT) began.

Figure I.5 History of RPI sonofusion resonators.
This overview of the sequence of resonators shows how various technical detail solutions have been developed in an effort to control problems like leaking, piston tilting, reproducibility of assembly, and in order to experiment with parameters like chamber fixation and component connection stiffnesses. The exemplars 1, 2, 4, 5, and 6 are (with the exception of the flanges) of the design tradition following [461, 505]. Exemplars 7 & 8 are based on the design of Cancelos [69].

Resonator Nº 1: The main glass cylinder with transducer has been furnished by Taleyarkhan's group and labelled by them as one of the three successful exemplars [392]. The aluminium flange at the top edge had been manufactured by the Purdue group, but glued to the glass cylinder only at RPI. The metal part (possibly tin) for bottom piston fixation and possibly the bottom piston as well stem from a second exemplar furnished by the Purdue group, which however had been broken during transport. This resonator is one of the two used for sonofusion trials [390] and has been described as one of the best performing exemplars [390]. It's pressure profile at 19.9 kHz is reprinted from [390] and

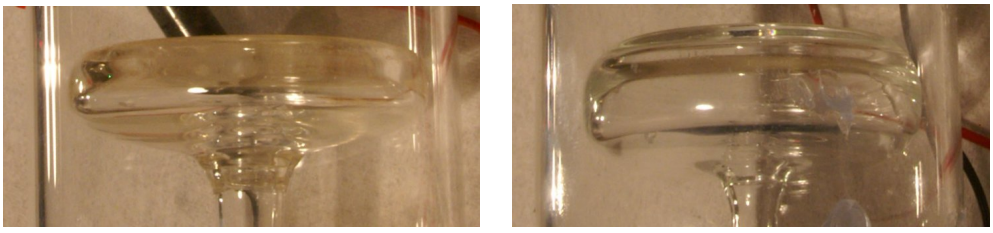


Figure I.6 Glass piston shape variations.
These are close-ups of the bottom pistons of resonators Nº 2 (left) and 5 (right). The first one has presumably been produced by flattening a sphere, because throughout the front and side wall the glass thickness is rather thin but with mostly radial variations including a slight thickening towards the centre of the front surface. It is one of the parts received from Taleyarkhan's Purdue lab [393]. The second one belongs to the set produced at RPI, where, in order to have a piston front wall of well-defined thickness, a different method has been used by the glassblower. Round discs have been cut out from a flat glass plate and fused to the end of a wide glass tube. Then the back end has been thinned and connected to the thin glass tube. These pistons have thicker glass walls and are much sturdier around the front and the side rim than the other ones. (Unfortunately, not all of these features can be seen in the pictures.)

can be seen in figure I.1. The Q-factor of that resonance has a value[7] of 340.

Resonator Nº 2: undocumented.

Resonator Nº 3: has been used by Frank Saglime for characterisation and preliminary cavitation experiments.

Resonator Nº 4: Being a copy of resonator Nº 1 and having been manufactured completely at RPI, this is the second resonator used for sonofusion trials [390].

Resonator Nº 5: This exemplar is the one whose acoustic properties have been thoroughly examined as part of this work. It was wholly manufactured at RPI.

Resonator Nº 6: This resonator represents a later trial of the RPI team. It has been tried to avoid the flanges, after none of the resonators with flanges had brought positive results. However, this exemplar has a different feature rendering it quite uncomparable to the other ones in terms of acoustic properties, namely the paper form filled with a massive cone of silicone invented here with the intention to be a more controllable fixation of the bottom piston.

Resonator Nº 7: It is of the design of Silvina Cancelos and has been examined by Frank Saglime and is the small version of the resonator shown in figure I.8 on the left. This was a resonator made of one single glass piece. The central liquid volume is laid out so the height of the central fluid volume between the piston surfaces matches its diameter, and the latter dimension has been kept the same ($\approx 6\,\mathrm{cm}$, as with the other resonators). The geometry has a horizontal symmetry plane only disturbed by the glass knobs for chamber suspension with wires. It also features enough nozzles for pressure control, draining, and refilling. The only problem is that it became damaged within the first minutes of experimentation [392] as a fatigue crack appeared in the glass near the top

[7] taken from figure 20. page 40 of [390]

rim, where the design shows a very high curvature of the glass wall, and where the material is under high bending stress.[8]

Resonator Nº 8: This is a twin of resonator Nº 7, sharing with it the exact same fluid volume geometry. However, instead of being all-glass, there are aluminium flange constructions for capping the top and bottom end. It is the small version of the resonator shown in figure I.8 on the right. A main reasoning behind the development of this design by Cancelos [69] is the better reproducible part machining and assembly with minimised tolerances, because no manual glassblowing is part of the manufacturing process.

Looking at that collection of hardware, several additional facts should be mentioned:

- The starting point for all the resonators built at RPI, all information input concerning desirable dimensions and efficient techniques, consisted in the parts of resonator Nº 1 and the annotated sketch in figure D.1, all supplied by Taleyarkhan and the Purdue group, accompanied with some explaining notes and a sonofusion experiment protocol. Geometry and construction variation details that had been tried out before by Taleyarkhan with his laboratory team at ORNL, who are said to have examined more than forty unsuccessful resonators [253], are largely unknown.

- Large differences in the sound pressure/frequency response and the achievable cavitation rate were observed by the RPI team, even between resonators of similar design and setup, and even between different runs with the same resonator at different times. This shows that reproducibility is a major problem.

- The sources of the performance variations are assumed to include:

  a) variations in transducer properties and vertical position,

  b) variations in temperature and thus the liquid's speed of sound,

  c) variations in resonator suspension (clamping, hanging, bolting, etc.),

  d) variations in the acetone filling level,

  e) variations in vertical piston positions,

  f) pistons being not well aligned and variations in tilting angle,

  g) variations in the shape of glass parts,

  h) variations in the acoustic coupling strength of silicone connections, bolted connections, squeezed O-ring connections,

  i) variations caused by different persons assembling and handling the hardware differently (tightness of bolts or clamps, positioning of flexible parts, silicone usage, glassblowing etc.),

[8] That the top and bottom rims, where the glass wraps around a 180° curvature, are points of high dynamic material stress, can be seen in Fig. 17 on page 42 in [69], which represents a FEM simulation snapshot obtained by Silvina Cancelos.

j) many more variations, e. g. glass cylinder not perfectly round, asymmetric epoxy layer thickness, material properties of cured epoxy and silicone, etc..

- The top heads mostly have only one nozzle connected to the inner volume, and this nozzle had been used for pressure control and continuous degassing. Continuous degassing leads to a loss of acetone through evaporation. If no port for topping off the liquid is available, then the liquid level cannot be kept constant. The addition of liquid requires the opening of the resonator.

- Unlike the ORNL sonofusion trials by Taleyarkhan et al., where the resonator head piece had always been connected to the rest by a flexible silicone bead, most RPI trial resonators (except Nº 3 & 6) involved the bolted flange constructions seen in the pictures. The idea and the first aluminium part (seen at the top rim of resonator 1) came from the Purdue group [393]. The primary purpose was to ease the process of refilling or topping off the liquid content. However, the introduction of the flanges severely impacts the acoustic behaviour (discussed in section O.3.3).

- The glass pistons represent a particularly sensitive aspect of the resonator design. Not only their positions inside the resonator influence it, but also their intrinsic acoustic properties determined by their shapes. These variations in shape have three main reasons. The first reason is the different piston diameters pointed out in the sketch by R. Taleyarkhan, shown in figure D.1. The second reason consists in the nature of manual glassblowing, where not all details of how gravity, surface tension, and viscosity gradients shape the glass are totally under control. The third reason, contributing to a quite substantial variation of proportions stems from what method a glassblower chooses when manufacturing them and what the craftsman is aiming at. Figure I.6 shows close-ups of the lower pistons of resonators 2 & 5. The piston of resonator 2 has probably been made by flattening a spherical glass bubble. In any case its maker has been aiming much more at minimising the thickness of the front plate and the side wall compared to the maker of the other one. The lower piston of resonator 5, like several other pistons made at RPI has been fabricated with the primary goal of making the piston's front surface perfectly flat and even and having a controlled thickness of the front plate. This has been achieved by using round discs cut out from industrially produced glass plates and fusing them to the rest. As a consequence, these pistons exhibit larger wall thicknesses over front and side rim which makes them a lot sturdier.

- Various ways of piston fixation methods can be seen in the pictures. These techniques reflect different intentions (i. e. focus on alignment, focus on rigidity, focus on weak coupling, focus on leak-tight connection to the vacuum system, etc.). This variation is surely one of the main contributions to making the resonators uncomparable among each other. The metal part holding the bottom piston of resonator Nº 1 is a prime example of a small implementation detail that can be assumed to influence the acoustic properties of the resonator substantially. If it is connected tightly to both glass parts, the main glass cylinder

and the piston tube, then these are the consequences: (a) stronger coupling of motion, (b) the glass wall in connection with the top surface of the metal part can only move up or down, but radial motion and rotation (varying the angle against the cylinder axis) is strongly restricted, (c) possibly increased damping.

- The RPI team became aware of several sensitive aspects of the early resonators and tried to improve the design: (a) The flanges are targeted at making the resonator properties before and after refilling more similar. (b) More nozzles in the later designs are aimed at abolishing the need for opening and closing. (c) The use of glass plate discs for the pistons is aimed at making the acoustic properties of the pistons more controllable. (d) The transition to symmetric resonators where the transducer sits in the middle and the height of the main liquid volume equals the diameter aims at the formation of a standing wave with exactly one pressure antinode in the centre and low sound pressure amplitudes near the walls. But all those improvement measures are connected to other property changes, which may entail substantial disadvantages. The flanges add weight, change the vibration behaviour, lead to increased damping. The RPI pistons are much heavier and sturdier. The symmetric resonator approach has led to a high-Q glass chamber that broke quickly and a very low-Q (as will be seen) resonator (№ 8), made of much solid material per unit liquid volume.

- There are numerous resonators to compare, but the comparisons are not very telling if two resonators differ in several aspects. Unfortunately, comparisons exploring the performance of single resonators under different setup conditions have not been documented.

- The vibration mode of the standard setup (resonators 1, 2, 4, 5, 6) has not been understood well. This can bee seen in the thesis of Saglime [390] where the mismatch between the mapped pressure field (figure 21, reprinted here as figure I.1) on the one hand and simulation results (figure 15) and the author's expectations on the other hand were not fully resolved.

- Taleyarkhan's Purdue group were also working on FEM simulations and tried to compute the acoustic fields that were matching measurement data and could be useful for acoustic cavitation. Results from FEM simulations conducted by Adam Butt, published in 2005 [66], and later by Jing Wang ([498], published 2007) show acoustic fields which encompass on the one hand only one pressure antinode and on the other hand exhibit large sound pressure amplitudes (comparable to the central antinode) near glass surfaces. From their group there were no publications describing or explaining an acoustic field such as that shown in Saglime's hydrophone mapping ([390], fig. 21) until 2009, when such an acoustic field was plotted in a publication from Jing Wang et al. ([496, 497], fig. 4.1).

The resonator design of Silvina Cancelos

Silvina Cancelos had been conducting research on the influence of ultrasound on the permeability of living tissues towards pharmaceutical substances. The chosen source of high-amplitude and high-frequency ultrasound was the implosion of cavitation bubbles in water. Cavitation inside acoustic resonators has the advantage of a well-defined spatial distribution of the cavitation events concentrating around the sound pressure amplitude. An acoustic chamber had to be designed and constructed to be able to produce a sufficiently high rate of cavitation in water and which was large enough to host the tissue sample holder without too much perturbation of the sound field. This project was conducted in close cooperation with the bubble fusion team. The SF team's experience was helpful for Cancelos when designing a new resonator with a large outlet for the insertion of the tissue sampleholder. On the other hand there was the chance to try out a new symmetric resonator geometry in conjunction with the transition to an assembly without manually formed glass parts, which promised to be valuable to the SF team.

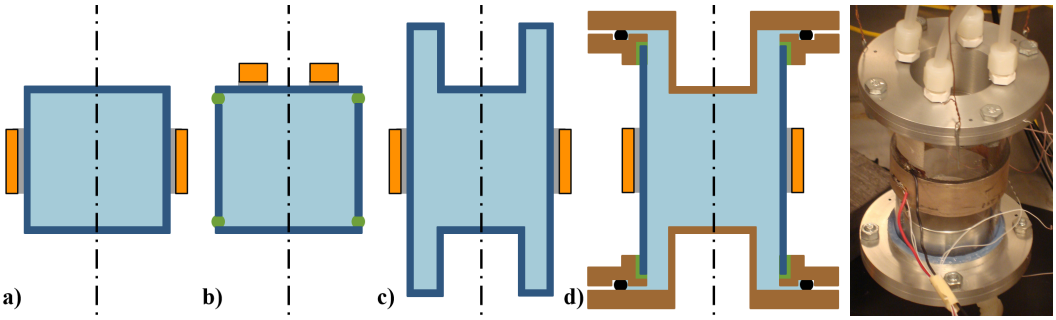


Figure I.7 Steps leading to the design of Cancelos.
A simple resonator geometry to be driven by a transducer in the form of a hollow cylinder is the geometry on the left. It can be designed to yield a desired frequency of the fundamental resonance by making the right choice of dimensions for a given velocity of sound in the working liquid. The diameter matches the height, and both match half a wavelength of the resonant sound wave. The walls are to become pressure nodes and displacement antinodes while the centre point of the inner volume should be the location of the displacement node and pressure antinode. As the flat end plates have to move inwards and outwards in phase with the cylinder wall, high bending stresses and fatigue are predicted at the cylinder edges. Vertical extensions of the side wall at both ends of the cylinder are the countermeasure leading to the geometry depicted in the centre. A resonator with exactly the same geometry of the inner volume, but built of an assembly of precision-machined parts is shown on the right.

After considering the design steps outlined in figure I.7 the resonators shown in figure I.8 were constructed. The one exemplar with aluminium flanges has been the main tool for the experimental work. It yielded a satisfying cavitation rate in water at 18 °C with a resonance frequency close to 14 kHz. The cavitation threshold had been observed driving the resonator at around 4 W dissipative power. The resonator had been used at a setting of 3 W for no cavitation and 6 W for cavitation. These values have been noted as corresponding to 252 and 286 kPa, respectively. The voltage amplitudes on the transducer seem to have been 23 and 40 V. The Q-factor of that resonator had been given[9] as $Q = 60$.

[9] In [69] the Q-factors from pressure signals were calculated based on the peak width at $\frac{1}{2}$ the peak height instead of $\frac{1}{\sqrt{2}}$ the peak height. Reinterpreting the data plotted in [69] in figure 37 with the formula given in table table J.2 (p. table 307) yields $Q = 98$.)

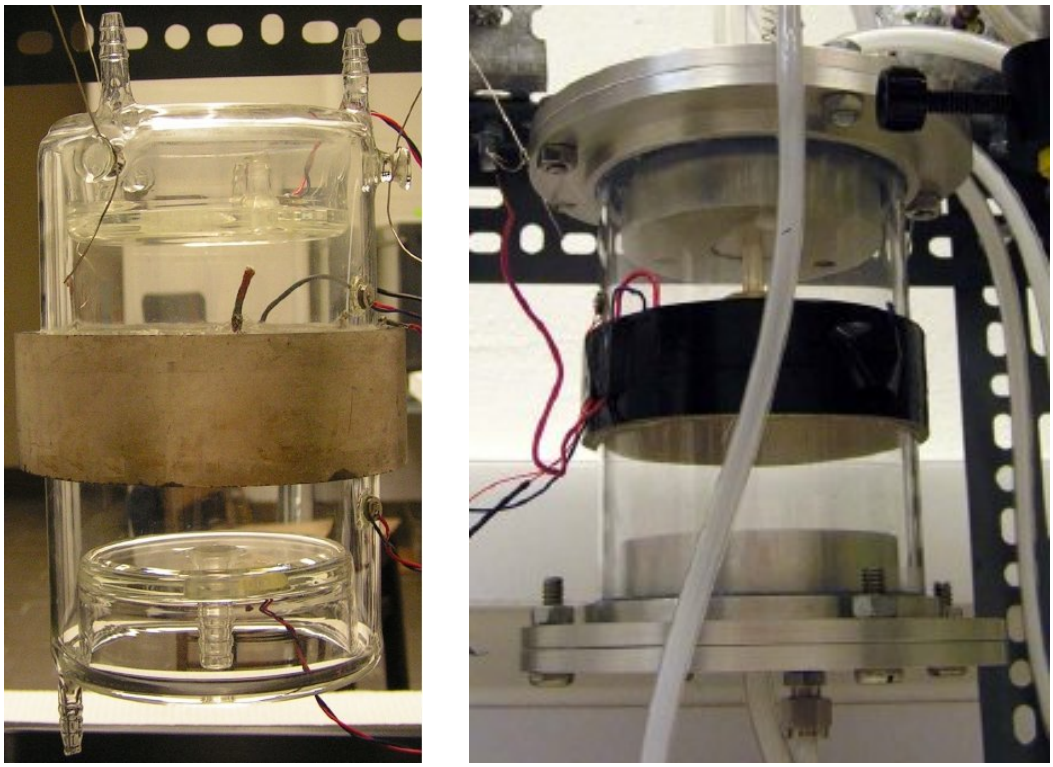


Figure I.8 Resonators employed by S. Cancelos.
The resonator on the left is made of one glass piece. It has nozzles for filling, draining, and pressure maintenance and an additional access on the central axis for the hydrophone. It has been used for simulation validation. The resonator depicted on the right features the same fluid volume geometry, but instead of being one solid piece of glass it is an assembly of a cut-out section of an industrially manufactured glass tube and four aluminium flange parts. Two flange rings are glued with silicone to the glass cylinder. Two more flange parts, the end caps, are fixed through bolts and sealed with rubber O-rings. One end plate has a central outlet of larger diameter allowing the introduction of either the hydrophone or the tissue sample holder. It can also be seen that Cancelos transitioned to suspending the resonators by wires. (Reprint with permission from S. Cancelos)

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
E	energy
k	Boltzmann constant
N	event count
Q	quality ("pointedness" of a resonance peak)
T	temperature
W	work, energy
w	weight coefficient
x	distance, spatial coordinate

List of Greek quantity symbols

Symbol	Description
η	nucleation efficiency
Φ	flux

List of particle symbols

Symbol	Particle
α	alpha particle consisting of $2p + 2n$
D	deuteron consisting of $p + n$
e^-	electron
γ	photon
p	proton
n	neutron
T	triton consisting of $p + 2n$

List of abbreviations

Abbreviation	Description
FE,FEM	finite element (method)
MCNP	Monte Carlo N-Particle (simulation code)
NYSERDA	New York State Energy Research and Development Authority
KIT	Karlsruhe Institute of Technology (Karlsruher Institut für Technologie)
ORNL	Oak Ridge National Laboratory
PDF	probability density function
PNG	pulsed neutron generator
RPI	Rensselaer Polytechnic Institute
SF	sonofusion
SRIM	Stopping and Range of Ions in Matter (simulation code)

Appendix J

Some basics of transducer analysis

J.1 Resonators as energy flow filters

A spring-mass system is a resonator. It can very often be approximated as a damped harmonic oscillator exhibiting its well known resonance behaviour. Figure J.1 shows a generic oscillator and its frequency response. If the motion of the mass is damped by friction then during stationary oscillation there will be a constant flow of energy from the excitation device over the spring to the mass, from where it is dissipated through the friction phenomenon. The amplitude plot (top right in fig. J.1) shows that for a frequency $\omega >> \omega_0$ far beyond the resonance frequency ω_0 the oscillation amplitude of the mass tends to zero, this is the case when the excitation signal oscillates so fast that the mass cannot follow any more because of its too high inertia in combination with too weak a spring. If there is no oscillation and hence no dissipation, then it also means that the energy transfer is quenched. That regime can be reached either by going to the limit of high working frequencies or via a high mass or a weak spring which both reduce ω_0. Thus, in the extreme cases energy flow from left to right through the spring in the sketch, i. e. from the excitation device to the oscillator is marginal.

Instead of thinking of friction as an energy sink we can also think of a second spring connecting the mass on the right to another system which can receive energy. But it is clear that energy can only flow into the receiver system if it reaches the centre mass first. Under the assumption that the excitation system on the left is very heavy and sturdy so it will not be disturbed by whatever is attached to its right port, it is the combination of the excitation frequency, the weight of the mass, and the stiffness of the spring which together determine how well energy can flow into the centre mass. The second spring's stiffness together with the internal stiffness- and mass-like properties of the receiver system on the right and its dissipation characteristics determine how well energy can continue to flow from the centre mass to the right. In technical terms, energy flow is maximised *if the impedances match.* The left diagram in figure J.2 shows how a generic transducer can be characterised as a network component by specifying transduction coefficients and terminal impedances.

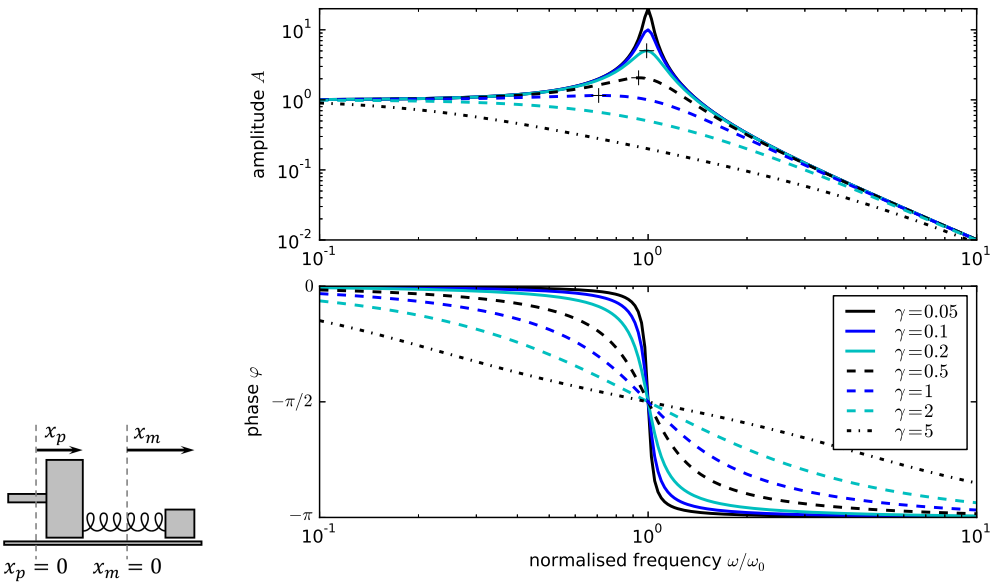


Figure J.1 The harmonic oscillator
The sketch on the left shows a simple form of a harmonic oscillator, a spring-mass system with a linear spring. In an oscillator formed by a spring-mass system the stored energy shifts between the two forms kinetic and potential energy. (There are however other oscillators with analogue characteristics, e. g. an electric circuit with an inductance and a capacitance, where the stored energy is used to create different electromagnetic fields inside a capacitor and around a coil and where the former component serves as a storage facility of potential energy and the latter one's effect is analogue to inertia.) The mass m in the above sketch is connected with the spring of stiffness k to the heavy piston on the left. The temporally changing position $x_\mathrm{p}(t)$ of the heavy piston is the excitation signal to the spring-mass system. The excitation piston can be imagined "to be very heavy" in order to reflect the condition that the spring force has no back-coupling influence on the excitation signal. The fact that the spring is linear allows the direct translation of the offset of the piston from the reference position into an additive excitation force $f_\mathrm{excitation}(t) = f_\mathrm{excitation}(x_\mathrm{p}(t)) = k \cdot x_\mathrm{p}(t)$ acting on the mass m and being independent of the mass' own position. The mass moves as a response to $f_\mathrm{excitation}$ and is also influenced by its own inertia. If the heavy piston comes to rest after some movement, then the mass will keep on oscillating at its resonance frequency until all the energy stored in the spring-mass system has been dissipated by friction. The plots on the right show the response of the mass to an excitation signal of stationary harmonic form (for the realistic case that damping by friction is present and under the particular assumption of velocity-proportional damping, see below in section K.5). The response itself is a stationary harmonic oscillation. The upper plot on the right shows the amplitude response and the lower one the phase response.

J.2 What is a transducer?

Often, and preferably when two different forms of energy are involved, such an energy-transferring network component is called a *transducer*. In case the component contains internal[1] spring-mass systems this will lead to a nontrivial frequency-dependence of its transfer function. The human ear is a good example of a transducer, consisting of the eardrum (spring), the ossicles (masses), and the liquid-filled inner ear (the cochlea, the receiver system). Its sole purpose is to make impedances

[1] Often the word "transducer" is used with a broader meaning of transforming any kind of signal or energy to any other kind. Such examples are: electrical thermometer, flow meter, light emitting diode, photo cell. These have no internal equivalent of a spring-mass system.

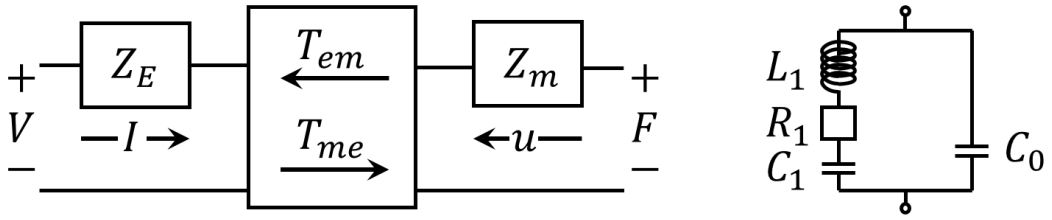


Figure J.2 Schematic of a transducer and exemplary equivalent circuit
The left plot gives a general symbolic representation of an electromechanic transducer, according to O. B. Wilson [516]. The left port is the electrical one (voltage) and the right port is the mechanical one (force). T_{em} is the transduction coefficient giving the electromotive force at the left terminal per unit of velocity on the right port, T_{me} represents the force induced in the mechanical port per unit of current going through the left terminal Z_E and Z_m are the input electrical and mechanical impedances observable with the opposite terminal blocked, which means clamped for the mechanical and open for the electrical one. Piezoelectric transducers can be represented by various equivalent electrical circuits which allow to approximate these impedances and transfer coefficients and to emulate the internal resonance pattern by adjusting the specifications of the equivalent electrical components. The diagram on the right shows a simple equivalent circuit for piezoelectric transducers. This circuit has been popularised by official IEEE standard documents [213, 214]. It is called a *Butterworth-Van Dyke (BVD) circuit* [24]. Here, C and C_0 are the *motional* and the *parasitic capacitance* [107]. The latter is also called *blocked* [516] or *parallel capacitance* [214]. R and L are the series resistance and inductance. The L-R-C branch alone is equivalent to a spring-mass system where L represents the inertia, $1/C$ the stiffness, and R the damping coefficient [107]. The meaning of the second branch with the parallel capacitance C_0 is the leak path existing for AC current as the transducer electrode surfaces and the supply cables form in fact a capacitor [107].

match. If the oval window of the inner ear was exposed directly to the external sound-carrying air, then we would hear almost nothing because the acoustic impedances between the air and the liquid filling the inner ear do not match. The displacement signal of the sound wave in the air is captured by the relatively large eardrum, and the collected force signal gets focused and transferred by the ossicles to the oval window with its much smaller surface. Here, the concentrated force signal can generate a displacement wave of much larger amplitude in the liquid of the cochlea as would be possible through a direct air-liquid interface where most of the sound energy would simply be reflected. A microphone is a similar transducer system. Its central part is a small membrane excited by the pressure waves coming through the air. But the pickup system is different: in a microphone the translational signal of the membrane is transformed into an electrical signal with the help of e. g. a magnet-coil system or a capacitor. In a magnet-coil system either the magnet or the coil has to be mounted flexibly. The elasticity of that mounting and the air membrane together with the masses of both parts determine the resonance behaviour of the system.[2]

Liquid-filled acoustic resonators are the subject of this project, and piezoelectric transducers in the form of radially polarised hollow cylinders are the drivers for the resonator design of interest. Layers of silver on the inside and outside surfaces of the hollow cylinder form the electrodes of the electrical port. They can be seen

[2]In a microphone, the resonance frequency lies high above the audible range. In the inverse system, a loudspeaker, the resonance frequency is below the working frequencies. The microphone is made to work in the regime $\omega \ll \omega_0$ where the pressure oscillations of the air appear slow so the internal mechanics of the microphone can track the $p(t)$ curve with fidelity even though the force signal is weak. A loudspeaker is made to work in the regime $\omega > \omega_0$ to $\omega \gg \omega_0$. It is strongly damped so it shows not much of an amplitude bulge around ω_0 and only a slowly decaying amplitude above. It is not such a big problem in that case because the excitation force can be easily made strong enough so the weak elastic bearing is of small influence.

as capacitor plates creating an approximately homogeneous electrical field between them which acts on the piezoceramic. With that electrode geometry the radially polarised piezo hollow cylinder can be forced to motions of radial contraction and expansion by applying a voltage across the electrodes. Subject to an alternating voltage signal, the hollow cylinder oscillates between contraction and expansion. When the ring's resonance frequency is matched, inertia and elasticity lead to an overshooting of the radial motion and a substantial amount of energy is stored in the transducer and periodically shifts between kinetic and potential energy. When the transducer is glued to an empty or filled glass vessel as in the resonator application, then this additional mechanical load modifies the resonance characteristics.

J.3 Impedance, admittance, and the BVD equivalent circuit

The resonance pattern of an electromechanical transducer can be measured from both sides, by gathering force, velocity, or displacement data from the mechanical side and by recording current and voltage signals on the electrical port. When the transducer is excited with a sinusoidal signal, e. g. a force $F(t) = F_0 \sin(\omega t)$ on the mechanical port or a voltage $U(t) = U_0 \sin(\omega t)$ on the electric port, then the transducer, if kept within the linear range of amplitudes, will settle into a harmonic oscillation so any other observable signal $s(t)$ will be of the sinusoidal form $s(t) = s_0 \sin(\omega t + \varphi_s)$. Each signal can be fully characterised by giving the two numbers s_0 and φ_s, the amplitude and the offset phase angle with respect to the excitation signal. These 2D data pairs can also be interpreted as points in the complex plane. Examining the transducer from the electrical side, the observables are voltage U and current I. Writing them not as temporal signals but as complex numbers one has

$$U = \operatorname{Re} U + i \operatorname{Im} U = U_0 \left(\cos(\varphi_U) + i \sin(\varphi_U)\right) \tag{J.1}$$

$$\text{and} \quad I = \operatorname{Re} I + i \operatorname{Im} I = I_0 \left(\cos(\varphi_I) + i \sin(\varphi_I)\right). \tag{J.2}$$

By convention, the writing of Ohm's law, $U = RI$, turns into $U = ZI$, yielding the definition of the impedance Z where there used to be the resistance R. Similarly, the conductance G turns into the admittance Y. We have

$$Z = R + iX \tag{J.3}$$

$$\text{and} \quad Y = G + iB = 1/Z. \tag{J.4}$$

This notation is helpful, because when describing many physical oscillating systems with complex numbers, their math rules work out in predicting reality, e. g. Kirchhoff's circuit laws can be applied. Where in the realm of steady flow network problems real quantities like resistance R or conductance $G = 1/R$ suffice for description, in the realm of oscillating flow networks (not only electrical engineering, but also e. g. acoustics) the quantities *reactance* X and *susceptance* B have to be added as imaginary parts, in order to solve questions of wave and energy transport.

Figure J.3 shows an exemplary electrical characterisation data set gained from an FE simulation of a freely oscillating piezo hollow cylinder. It can be seen that there

are two characteristic frequencies involved in the resonance pattern, the resonance itself around $f_1 \approx 16.12\,\mathrm{kHz}$ and the *antiresonance* $f_2 \approx 16.67\,\mathrm{kHz}$. The resonance ($f_1$) is marked by peaks of $|Y|$ and $\mathrm{Re}(Y)$, whereas the antiresonance has peaks in$|Z|$ and $\mathrm{Re}(Z)$. In the vicinities of f_1 and f_2, there are also points marked by a zero phase angle of Y and Z. Table J.1 lists all the characteristic frequencies. In the limit of vanishing damping f_m, f_s, and f_r fall into one place f_1 and $f_a = f_p = f_n$ into f_2, but in the case of a lossy transducer the two triplets split up as can be seen in the plot.

Table J.1 The characteristic frequencies of electromechanical transducers

label	description
$f_1 \to (f_m, f_s, f_r)$	resonance (with $f_m \le f_s \le f_r$)
$f_2 \to (f_n, f_p, f_a)$	antiresonance (with $f_a \le f_p \le f_n$)
f_m	maximum $\lvert Y\rvert$
f_s	maximum $\mathrm{Re}(Y)$, *motional resonance, series resonance*
f_r	$\mathrm{Im}(Y) = 0$, *electrical resonance*
f_n	maximum $\lvert Z\rvert$
f_p	maximum $\mathrm{Re}(Z)$, *parallel resonance*
f_a	$\mathrm{Im}(X) = 0$, *electrical antiresonance*
order:	$f_m \le f_s \le f_r < f_a \le f_p \le f_n$
f_{mB},f_{nB},f_{mX},f_{nX}	corresponds to max B, min B, max X, min X
order:	$f_{mB} < f_s < f_{nB}$ and $f_{mX} < f_p < f_{nX}$

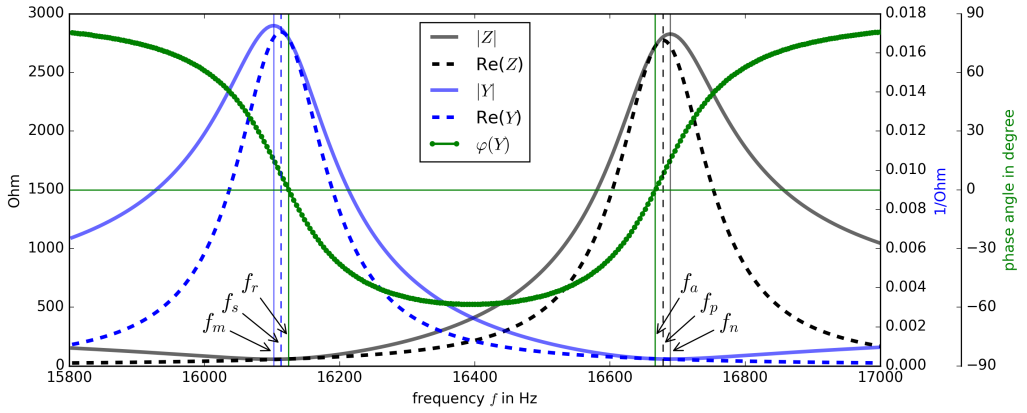


Figure J.3 Transducer characterisation from the electrical terminal.
This plot shows the electrical characterisation data gained from a forced harmonic FEM simulation of a freely oscillating piezo hollow cylinder. On the electrical terminal of an electromechanical transducer, harmonic current and voltage data can be recorded as complex numbers, based on amplitudes and phase offsets. The impedance $Z(\omega)$ and the admittance $Y(\omega)$ can be computed: $Z = U/I$ and $Y = 1/Z$.

It is often preferable to condense much of the measured data contained in a plot like figure J.3 into a much simpler form using plots in the complex number plane where the complex admittance and impedance describe circles, as shown in figure J.4.

Taking measured admittance and impedance circles as a starting point and the

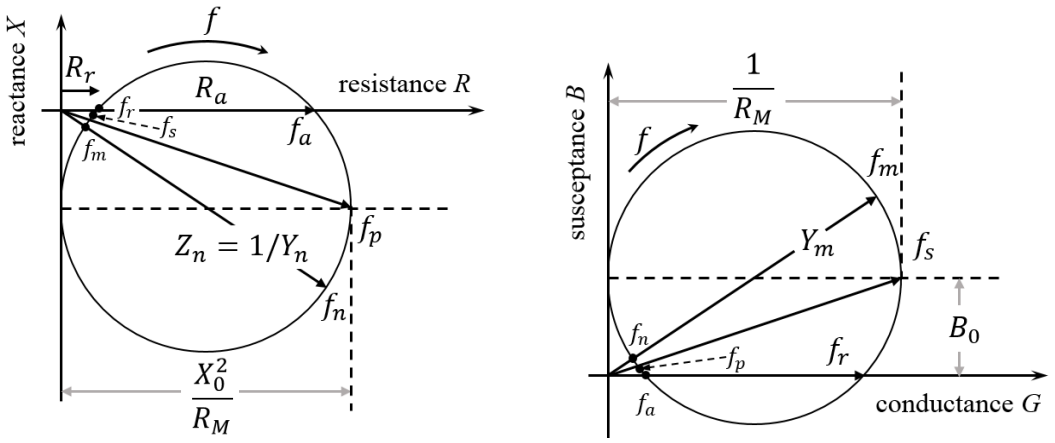


Figure J.4 Impedance and admittance circles: Tracking a transducer's impedance Z in the R-X plane, or its admittance Y in the G-B plane while varying ω as a parameter in the vicinity of a resonance will lead to circle-shaped plots. These impedance or admittance circles are commonly used because they offer an easy way to determine various quantities of physical meaning. For example, circle sizes and the frequency spread tell about the damping and the vertical offset of the circle centre from the real axis tells about $C_0 = B_s/\omega_s$. It also becomes obvious that increasing the vertical offset enlarges the split between f_m, f_s, and f_r (or f_a, f_p, and f_n), but it does not affect the split between f_{mB} and f_{nB} (or f_{mX} and f_{nX}) and thus Q_m.

Butterworth-Van Dyke (BVD) equivalent circuit (figure J.2) as the model basis, it can be asked which values of R, L, C, and C_0 represent the case, in which the admittance

$$Y(\omega) = i\omega C_0 + \frac{1}{R + i\omega L + 1/i\omega C} \tag{J.5}$$

of the circuit and its impedance $Z = 1/Y$ best resemble the measured plots. Furthermore, instead of requiring a general fitting procedure, the properties of the BVD model allow to use a set of shortcut equations to calculate various quantities suitable for characterising a given piezoelectric transducer. Table J.2 lists these formulae, compiled from [213, 214, 516] and [107][3]. Adopting the nomenclature of [214], f_1 is the collective term for the resonance splitting up into (f_m, f_s, f_r) and f_2 denotes the antiresonance (f_a, f_p, f_n). DeAngelis & Schulze [107] use the terms f_1 and f_2 for the top and the bottom of the admittance circle. These frequencies are labelled here as f_{mB}, the frequency of maximum susceptance B and f_{nB} the frequency of minimum B.

The meaning of the series resonance f_s and the parallel resonance f_p can be explained with the BVD circuit (figure J.2) in mind. At the series resonance the current going through the motional branch is about the same as the the driving current $I_{mo} \approx I_{dr}$ and the current passing the parallel branch I_p is marginal. By contrast, at the parallel resonance f_p the motional and the parallel branch are excited to oscillate against each other, $I_p \approx -I_{mo}$, and both these currents drastically exceed the currents leaving and entering the excitation device $I_{mo} \gg I_{dr}$.

[3] who compiled formulae from [426, 516]

Table J.2 Useful formulae for transducer analysis

formula	ref.	description
$k = \sqrt{1 - \frac{f_s^2}{f_p^2}} = \sqrt{\frac{C}{C_0+C}}$	[107, 214, 516]	electromechanical coupling coefficient (Paraphrasing Wilson, k^2 is, in a sense, like a transduction efficiency, but more precisely, it is a characteristic of the mechanism with all dissipative processes ignored.)
$Q = Q_m = \frac{f_s}{f_{nB} - f_{mB}} = \frac{\sqrt{L/C}}{R}$	[107]	quality factor Q or mechanical Q (higher means less losses)
$Q_e = \frac{B_s}{G_{\max}}$	[107]	electrical Q (lower means less dielectric losses), a measure of the vertical offset of the Y-circle from the real axis
$Q_e = \frac{1}{\tan \delta_e} = \frac{\varepsilon'}{\varepsilon''}$	[360]	where $\varepsilon = \varepsilon' + i\varepsilon''$ is the material's clamped dielectric permittivity (for Q_e, higher means less dielectric losses, in conflict with the above)
$Q_e Q_m = \frac{C_0}{C} = \frac{1-k^2}{k^2} = r$	[107]	Q-product is the same as the capacitance ratio
$R = \frac{1}{G_{\max}}$	[107]	motional resistance, represents the mechanical dissipation
$L = \frac{Q_m R}{\omega_s} = \frac{1}{C\omega_s^2}$	[107, 213, 214]	motional inductance
$C = \frac{1}{Q_m R \omega_s}$	[107, 214]	motional capacitance
$C_0 = \frac{f_r^2}{f_a^2 - f_r^2} C \approx \frac{B_s}{\omega_s}$	[107, 213]	parallel capacitance, also called parasitic capacitance, it is a result of the electrode surface geometry (and sometimes cables as well), and it affects the ability of the electrical driver unit to deliver energy to the transducer
$\Gamma = C\, d/A$	[214]	motional capacitance constant (d: linear dimension $\|\vec{E}$, A: electrode surface area)
$B_s = \omega_s C_0$	[107]	susceptance of admittance circle centre (susceptance at the series resonance f_s)
$r = \frac{C_0}{C}$	[107, 213]	capacitance ratio
$M = \frac{Q_m}{r} = \frac{1}{\omega_s C_0 R} = \frac{k^2 Q_m}{1-k^2}$	[213, 214]	figure of merit
$\delta = \omega C_0 R$	[213]	normalised damping factor
$\Omega = \frac{f^2 - f_s^2}{f_p^2 - f_s^2}$	[213]	normalised frequency factor
$Z = \frac{1}{Y} = \frac{i}{\omega C_0} \frac{\Omega - i\delta}{1 - \Omega + i\delta}$	[213]	equation for impedance $Z(\omega)$ or admittance $Y(\omega)$
$Q = \frac{f_{\text{peak}}}{f_{\text{hdp2}} - f_{\text{hdp1}}}$		generic Q-factor of a resonant system where $f_{\text{hdp2}} - f_{\text{hdp1}}$ gives the peak width at half dissipation power which corresponds to a reduction by a factor $1/\sqrt{2}$ for many signal amplitudes

J.4 Alternative equivalent circuits

Many alternative equivalent circuits have been proposed for modelling electromechanical transducers [13, 23]. It has been pointed out by Martin [288] that the BVD circuit has the shortcoming of representing only one type of energy dissipation mechanism by its incorporation of a resistor R, whereas real piezoelectric transduc-

ers generally had to be assumed to exhibit three types of dissipation mechanisms, namely elastic, dielectric, and piezoelectric losses.[4] Sherrit et al. argue [417, 418] in favour of using the complex circuit model. It looks exactly like the BVD circuit, except that on the one hand the resistor is missing and on the other hand the remaining three components (L, C, C_0) are given complex values instead of purely real ones. Thus, three different damping mechanisms are at work. This also means going from four to six determining parameters. Moreover, they point out that the fact that an ab initio description of a single resonance of a simple piezoelectric geometry (their example is a plate) does in fact need six parameters to be fully determined, is another feature in favour of the complex circuit model. The only disadvantages of that model are that it has been discussed less in literature, and that fitting it to measurement data is not as straightforward as in the BVD case, because the frequency response plots in the Y and Z plane do not any more represent exact circles.

J.5 Lossy transducers and equivalent circuits

Admittance and impedance plots can be used to compare the outcomes of finite element models of a transducer with calculations based on equivalent circuit models. The comparison below will be made for an FEM simulation of the transducer as used here in the models of the West-Howlett-style SF resonators. The Butterworth-Van Dyke (BVD) and the complex circuit model are the two considered equivalent circuits, they are shown in figure J.5. The BVD model can represent mechanical, but not dielectric losses. By contrast, the complex circuit model is able to represent dielectric losses through the imaginary part of the parallel capacitance.

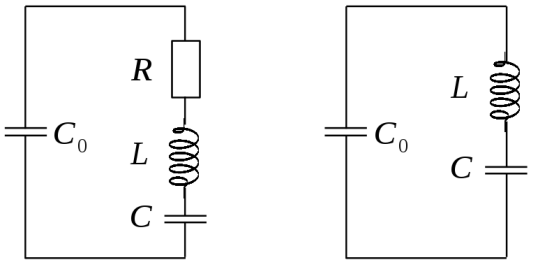


Figure J.5 The BVD and the complex equivalent circuit
The BVD equivalent circuit (left) is determined by four real-valued scalar parameters, R, L, C, and C_0. Energy dissipation takes only place in the resistor. The complex equivalent circuit is determined by six parameters forming the three complex numbers L, C, and C_0. In a harmonic analysis the imaginary parts of all three components model energy dissipation.

The plots of figures J.6 to J.8 represent results of an FEM simulation of the free transducer with ANSYS Classic® using `plane223` elements. Each shows several repetitions of the frequency sweep from 15.8 to 17 kHz in steps of 4 Hz under variation of a different loss mechanism. The FE model is the default 2D-axis-symmetric transducer geometry (cross section 3 mm × 25 mm and inner radius of 32.5 mm) with

[4] Going out from the mathematical description, in principle each entry of each material-describing tensor (see below in section K.4) can have its own imaginary part (i. e. loss tangent, see section K.5).

the properties of the C5800 material (label "C58b" in table Q.1). Additionally, a parallel (or "parasitic") capacitance modelled by a `circu94` element and connecting both electrodes has been added. The base values taken by the parameters if not varied are $\tan\delta_{\text{mech}} = 9.09 \times 10^{-4}$, $\tan\delta_{\text{diel}} = 4 \times 10^{-3}$, and $C_0 = 1 \times 10^{-20}\,\text{F}$. It can be seen in figure J.6 that increased mechanical losses lead to an even shrinking of both the Y- as well as the Z-circle and a reduction of both the admittance peak at the resonance ($f_r \approx 16.1\,\text{kHz}$) as well as the impedance peak (dip in $|Y|$) at the antiresonance ($f_a \approx 16.7\,\text{kHz}$). The dielectric losses act very differently (figure J.7): only the antiresonance peak loses its sharpness, while the admittance circle merely shifts along the real axis in the positive direction by a tiny margin. Lastly, the variation of the parallel capacitance (figure J.8) also shrinks the Z-circle and moves the Y-circle. But this time the move is up the imaginary axis, and the position of the antiresonance on the frequency axis is affected too.

The BVD model yields the admittance formula

$$Y(\omega) = i\omega C_0 + \frac{1}{R + i\omega L + 1/i\omega C} \qquad \text{with} \quad R, L, C, C_0 \in \mathbb{R}. \tag{J.6}$$

Taking one of the datasets generated by the FE simulation (the one of figure J.6 with $\tan\delta_{\text{mech}} = 0.005$), the parameters of the BVD model can be fitted to it by the routine described in appendix O.2.2, i. e. by fitting the Y-circle and incorporating the knowledge about f_a. This leads to the values $R = 11.7\,\Omega$, $L = 57.6\,\text{mH}$, $C = 1.69\,\text{nF}$, and $C_0 = 23.8\,\text{nF}$. Evaluating equation J.6 with these default values is the basis for figures J.9 and J.10 where the consequences of variations of R and C_0 are shown. These two cases correspond to the variations of $\tan\delta_{\text{mech}}$ and C_0 in the FE model, but the BVD model is incapable of representing dielectric damping.

Figures J.11 to J.13 show the behaviour of the the complex circuit model. The frequency response of the admittance is

$$Y(\omega) = i\omega C_0 + \frac{1}{i\omega L + 1/i\omega C} \qquad \text{with} \quad L, C, C_0 \in \mathbb{C} \tag{J.7}$$

for the complex circuit model. Variations of the imaginary parts of L and C have the same effect as varying R in the BVD model. The default values for $L = L' + iL'' = L'(1 - i\tan\delta_L)$, $C = C'(1 - i\tan\delta_C)$, and $C_0 = C_0'(1 - i\tan\delta_{C0})$ are $L' = 57.6\,\text{mH}$, $C' = 1.69\,\text{nF}$, $C_0' = 23.8\,\text{nF}$, and the value of 0.001 for each of the loss angles δ_L, δ_C, and δ_{C0}. Variation of δ_L and δ_C leads to indistinguishable plots, but the numbers are slightly different.

Thus, one could say that adding the real part of the parallel capacitance to the FE model and the imaginary part of it to the BVD model is bridging the gap between the two.

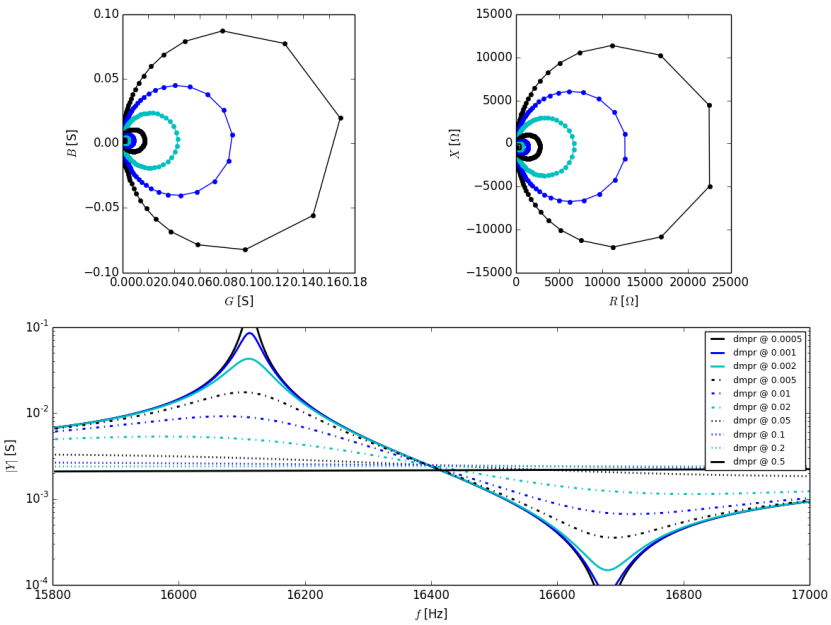


Figure J.6 Transducer FE model – varying mechanical losses.
This plot shows the effect of varying the mechanical losses using the APDL command `mp,dmpr` for entering the corresponding loss tangent or damping ratio, which can be thought of as adding a velocity-proportional damping matrix based on the stiffness matrix. In the harmonic case, it is in fact just the addition of an imaginary part to the stiffness matrix.

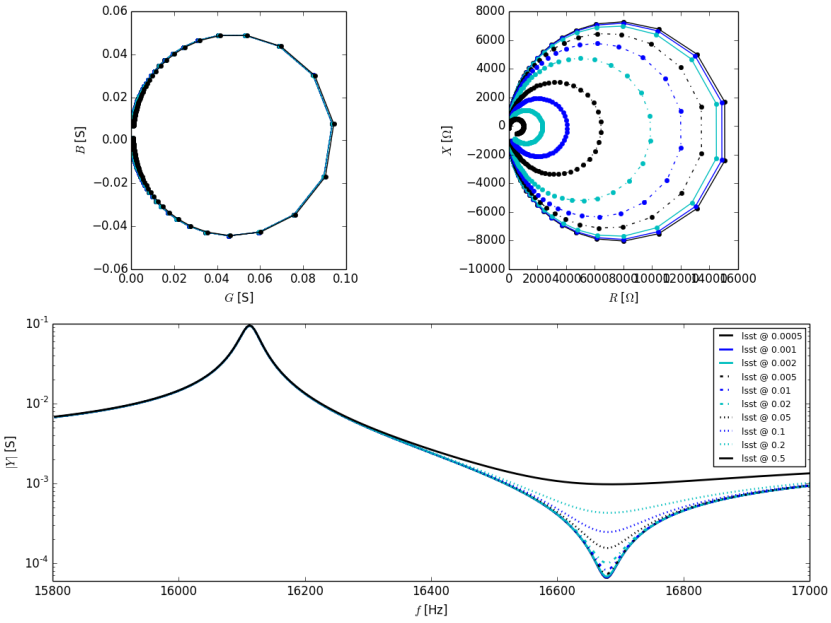


Figure J.7 Transducer FE model – varying dielectric losses.
Here, the effect of dielectric damping is shown, which consists of adding an imaginary part to the dielectric matrix by specifying another loss tangent via the command `mp,lsst`.

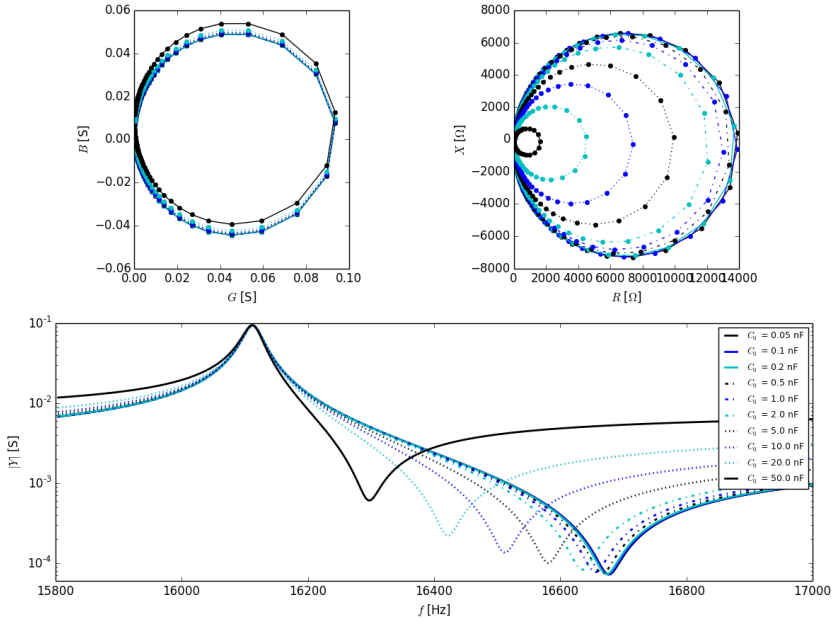


Figure J.8 Transducer FE model – varying the parallel capacitance
In order to generate these plots, a parallel capacitance has been explicitly added to the transducer model. It is a `circu94` element and connects the master nodes of both electrodes. In that case each electrode has to be a set of nodes coupled in their voltage DOF, so that the charge (the reaction force) of the whole electrode (the sum of all nodes) can be accessed through the master node [8]. The effects of an increased parallel capacitance C_0 are (a) shrinking of the impedance circle, (b) an upwards shift of the admittance circle (in agreement with the BVD model, and hence the formulae in table J.2), and (c) a shift of the antiresonance towards the resonance and thus a diminished coupling coefficient [353].

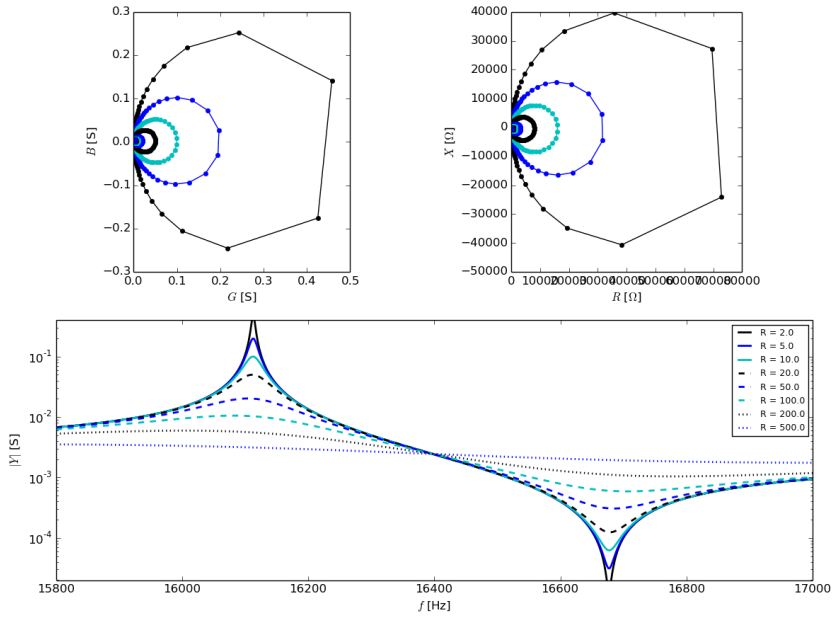


Figure J.9 BVD equivalent circuit – varying R
These plots represent the response of a BVD equivalent circuit, i. e. the evaluation of equation J.6, under variation of the resistance R.

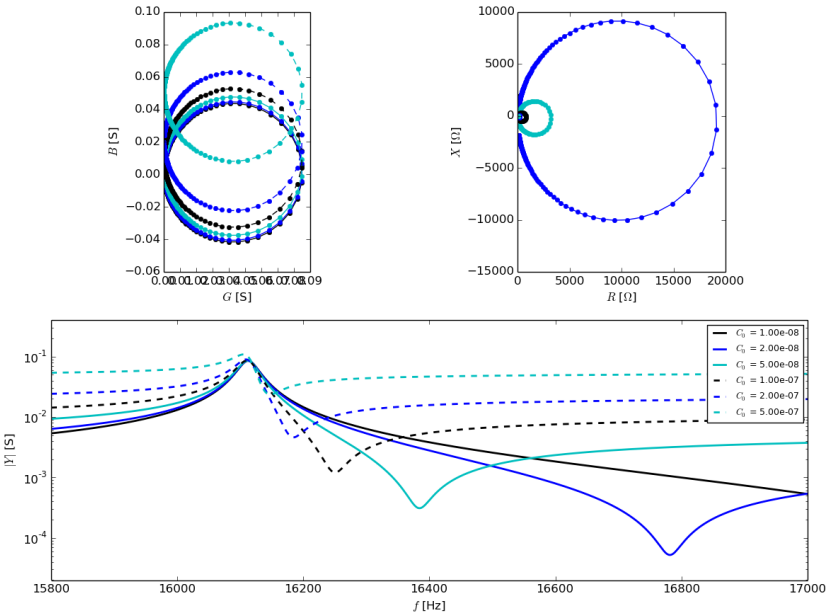


Figure J.10 BVD equivalent circuit – varying C_0
It can be seen that the increasingly large steps of the antiresonance towards higher frequencies when reducing C_0 reflects a different characteristic as given by the FE model in figure J.8.

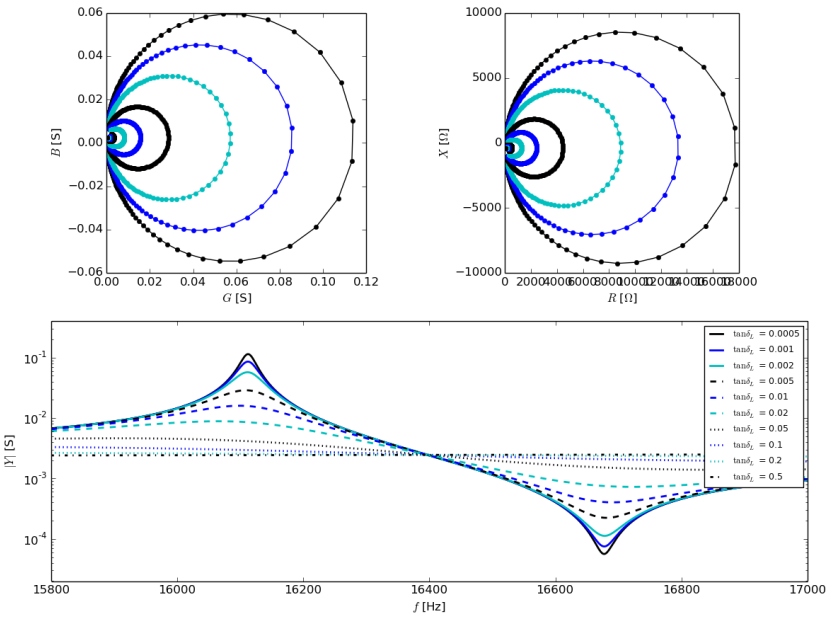


Figure J.11 Complex circuit model – varying the imaginary inductance
The inductance L must be set to $L = L'(1 - i \tan \delta_L)$ with a negative sign in front of the imaginary part for equation J.7 to produce reasonable values. A corresponding picture for the effect of $\tan \delta_C$ is not shown, because it would look exactly the same. The comparison shows that these two damping terms act like the R in the BVD model.

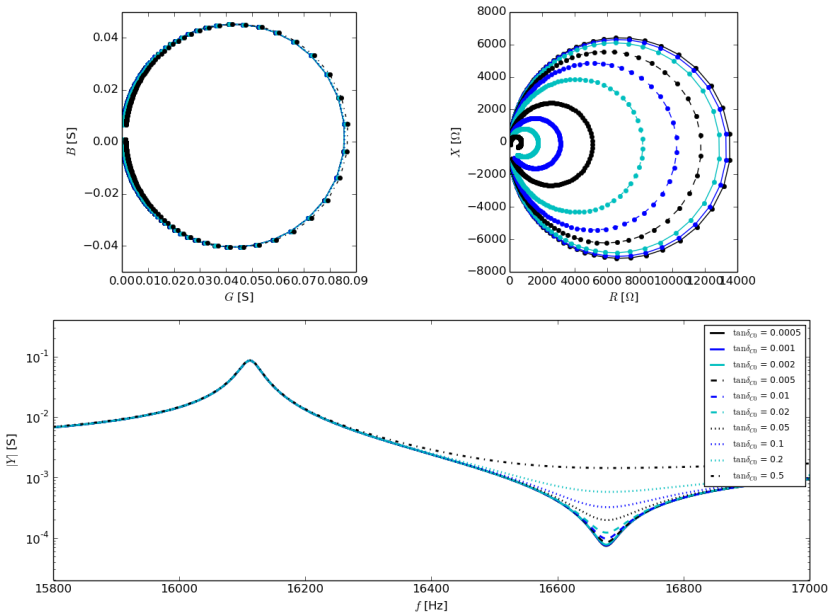


Figure J.12 Complex circuit model – varying the imaginary parallel capacitance
The loss ratio of $C_0 = C_0'(1 - i \tan \delta_{C0})$ is being varied. The impedance circle shrinks while the admittance circle keeps its size and only slightly shifts sideways. The location of the antiresonance on the frequency axis is not affected.

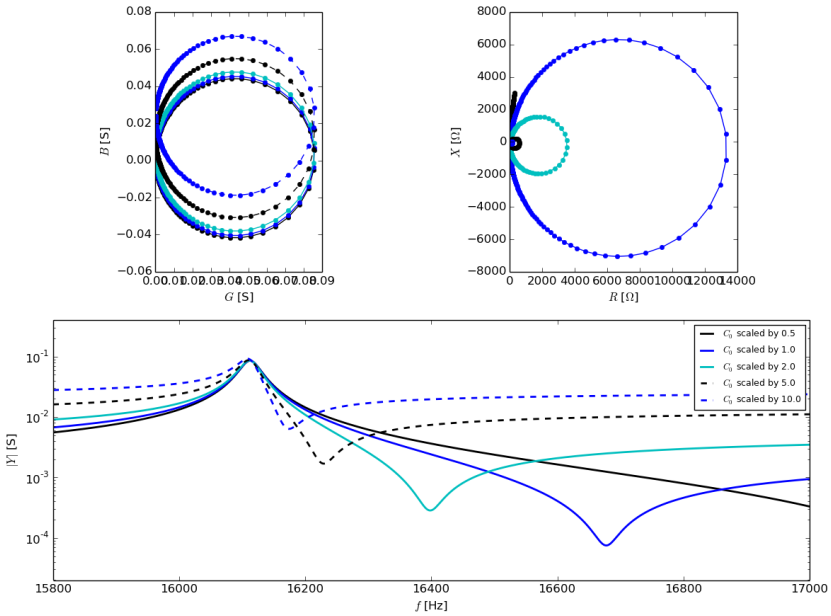


Figure J.13 Complex circuit model – varying the parallel capacitance

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
A	amplitude; surface area
B	susceptance
C	capacity
$\mathbb{C}$	complex numbers
E	energy
$\vec{E}, E$	electric field
F	force
f	frequency
G	conductance
I	current
i	unit value of imaginary numbers
k	electromechanical coupling coefficient
k	spring constant, stiffness
L	inductance
M	figure of merit
m	mass
p	pressure
Q	quality ("pointedness" of a resonance peak)
R	resistance
$\mathbb{R}$	real numbers
r	capacitance ratio
s	signal, obersable
t	time
T	transmissivity
U	voltage
X	reactance
x	distance, spatial coordinate
Y	admittance
Z	impedance

List of Greek quantity symbols

Symbol	Description
Γ	motional capacitance constant
γ	damping coefficient/factor
δ	normalised damping factor
ε	dielectric constant
ϕ	phase angle
Ω	normalised frequency factor
ω	angular frequency

List of abbreviations

Abbreviation	Description
APDL	Ansys Parametric Design Language
BVD	Butterworth-Van Dyke
DOF	degree of freedom
FE,FEM	finite element (method)
IEEE	Institute of Electrical and Electronics Engineers
SF	sonofusion

Appendix K

FE modelling of a piezo-driven acoustic resonator

Finite element methods (FEM) are used for finding solutions to systems of partial differential equations (PDE). The FEM concept has three main ingredients: (a) spatial discretisation into finite elements with own local coordinate systems that may change their relation to the global coordinate system, (b) simple trial functions defined on each finite element, common to a class of finite elements, defined in an element's coordinate system, and (c) transformation of the PDE system into a system of ordinary differential equations (ODE) with the help of a Galerkin projection onto the FE trial functions. At the core of the whole procedure are theorems of variational calculus. The conceptual closeness of the variational calculus forming the foundation of FE methods with the theorems behind other methods like the Lagrangian or Hamiltonian descriptions of classical mechanics is very well described in [36].

The FE simulations have been used in this work with the purpose of conducting so-called forced harmonic analyses. That means the stationary oscillatory response of the system is computed for a given harmonic excitation signal. For the SF resonator simulation the excitation signal is a sinusoidal voltage applied to the transducer's electrical terminal, and the computed oscillatory response data consists in the amplitude and phase offset values of quantities like charge, displacement, stress, and pressure at the nodes of the discretised geometry. The commercial FE software package ANSYS® has been used in this project because of its ability to model the electromechanical coupling in piezoelectric materials and the coupling of fluid and structure in acoustic simulations. The theoretical background of the FE method shall be sketched very briefly here. The details of how acoustic resonators were modelled with ANSYS in this work is documented in appendix Q and the associated appendices.

K.1 FE model generation for the acoustic domain

The lossless propagation of compression waves in a linearily elastic medium, is described by the wave equation

$$\frac{1}{c^2}\ddot{p} - \nabla^2 p = 0 \tag{K.1}$$

where the scalar field $p = p(\vec{x}, t)$ is the pressure and the constant c the speed of sound. c is related to the material's bulk modulus K and its undisturbed density ϱ_0 by $c^2 = \frac{K}{\varrho_0}$. Constant terms in p play no role in the equation, for there are only second derivatives present. Hence, it does not matter whether p is the label for the absolute pressure or the deviation from the mean pressure (the latter case will be taken as the working interpretation from now on). In a forced harmonic analysis one is only interested in the stationary vibration response of an acoustic domain (a volume Ω) to a stationary sinusoidal excitation signal (a pressure and/or displacement oscillation on a part of the domain boundary $\partial\Omega$). Therefore, the ansatz $p(\vec{x}, t) = P(\vec{x})e^{i\omega t}$ can be made where P is a complex number, and where the field $P(\vec{x})$ gives the amplitude and phase shift of the pressure oscillation in each point of the domain. With that ansatz the wave equation turns into the Helmholtz equation

$$\frac{\omega^2}{c^2} P + \nabla^2 P = 0. \tag{K.2}$$

The weak form of the statement can be made after multiplication with a test function v and integration over the domain:

$$\frac{\omega^2}{c^2} \int v P \,\mathrm{d}\Omega + \int v \nabla^2 P \,\mathrm{d}\Omega = 0 \quad \forall v \in \hat{V}. \tag{K.3}$$

Here, $\hat{V}$ is the space of all allowable test functions. Using the identity

$$\int_\Omega \nabla f \cdot \nabla g \,\mathrm{d}\Omega = \int_\Gamma f \nabla g \cdot \hat{n} \,\mathrm{d}\Gamma - \int_\Omega f \nabla^2 g \,\mathrm{d}\Omega,$$

derived in appendix L or [539], equation K.3 can be rewritten as

$$\frac{\omega^2}{c^2} \int v P \,\mathrm{d}\Omega - \int \nabla v \nabla P \,\mathrm{d}\Omega + \int_\Gamma v \nabla P \cdot \hat{n} \,\mathrm{d}\Gamma = 0 \quad \forall v \in \hat{V}. \tag{K.4}$$

It can be seen that the order of derivatives of P has been reduced from two to one.

No specifications have been made so far for the set of test functions $v \in \hat{V}$. The step from equation K.2 to K.3 can of course be made with any kind of function v, but useful mathematical advantages can be gained by choosing a linearly independent set of functions with special properties. The formulation implemented in ANSYS is based on equation K.4 and two more ingredients:

- The FE approximation: expressing P as the sum of interpolation functions, the *shape functions*.
- Inserting as the test functions v what are called *virtual displacements* δP of the solution P.

As an illustration of the first ingredient, some sample shape functions and decompositions based on them are illustrated in figures K.1 and K.2. With the concepts of virtual displacement and virtual work as the second ingredient, there arise further conditions for the approach to be physically meaningful. This results in the requirement that the ansatz for δP must be the same as the one for P.

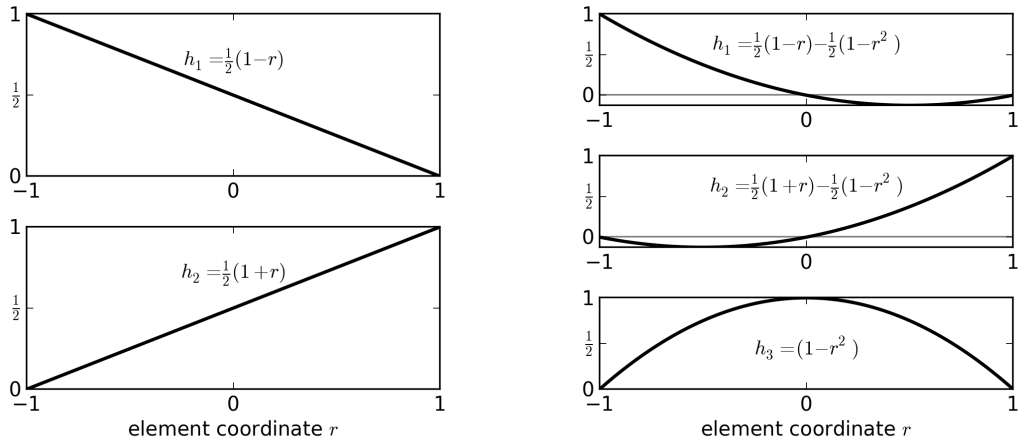


Figure K.1 Element shape functions for a one-dimensional finite element:
The left hand diagrams show linear shape functions for a two-node element with node 1 at $r = -1$ and node 2 at $r = 1$. The right hand plots show quadratic shape functions for a three-node element with node 1 at $r = -1$, node 2 at $r = 1$, and node 3 at $r = 0$. In both cases, the sum over the whole number of an element's shape functions $\sum_k h_k(r) = 1$ along the whole interval. The node and shape function numbers are associated so $h_i(r_i) = 1$ and $h_i(r_j) = 0$ for $i \neq j$, were r_j is the location of node j. (The presented formulae are the ones from chapter 5.3 of [31].)

There is a common notation for systems with many degrees of freedom (DOF) where vectors (think: lists of quantities) are written with curled brackets, e. g. $\{x\}$, and matrices connecting these vectors with square brackets like $[M]$. A list $\{x\}$ can be defined to contain the data of several conventional position vectors, like[1] $\{x\} := (x_a, y_a, z_a, x_b, y_b, z_b)^T$ or DOFs of different types like positions and angles can even be mixed. This makes the following shorthand writings for the FE approximation of the pressure solution possible:

$$P = \{N\}^T\{P_e\} \qquad \text{and} \qquad \frac{\partial^2 P}{\partial t^2} = \{N\}^T\{\ddot{P}_e\},$$

where $\{N\}$ is the list of element shape functions for the pressure and $\{P_e\}$ is the list of nodal pressure values. In this style the implemented FE system for acoustic problems is given in [8] as

$$\left(-\omega^2[M_e^P] + [K_e^P]\right)\{P_e\} + \varrho_0[R_e]\{\ddot{u}_e\} = \{0\} \tag{K.5}$$

with some additional shorthands

$$\begin{aligned}
[M_e^P] &= \frac{1}{c^2}\int_\Omega \{N\}\{N\}^T \,\mathrm{d}\Omega, \\
[K_e^P] &= \int_\Omega [B]^T[B] \,\mathrm{d}\Omega, \\
\varrho_0[R_e] &= \int_\Gamma \{N\}\{n\}^T\{N'\}^T \,\mathrm{d}\Gamma.
\end{aligned}$$

[1] Or, if it is more convenient, the listed quantities can represent vectors in different coordinate systems, like $\{x\} := (x_a, y_a, z_a, x'_b, y'_b, z'_b)^T$; FE systems can always be given either in the global or in the element coordinate system.

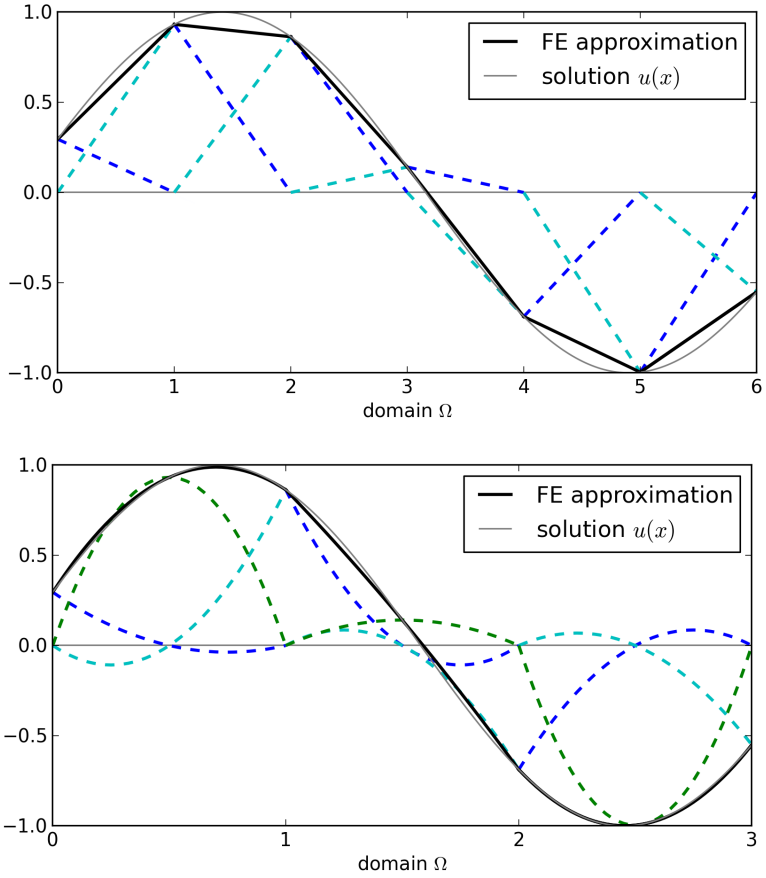


Figure K.2 Decomposition into shape functions:
These diagrams show the decomposition of a problem solution $u(x)$ into *shape functions* (or *hat functions* in the linear case). Be $u(x)$ a solution of the examined PDE in the domain Ω, shown as the smooth function in grey. The one-dimensional domain Ω has been discretised into an equidistant grid of six (three) finite elements with two (three) nodes each in the upper (lower) plot. Sums of the two (three) linear functions shown on the left (right) in figure K.1 are being used to represent $u(x)$ within each finite element which yields the piecewise linear (quadratic) interpolation function shown as solid black line. Within each element the approximation is made by the sum of h_1 (in dashed blue) and h_2 (dashed cyan) (and h_3 (dashed green)) with optimal weights.
In the upper plot one can shift the view and in thought concatenate the h_2 of the k^{th} finite element with the h_1 of the $(k+1)^{\text{th}}$ one to form little triangles or hats. Now, the black curve can be seen as the sum of triangular *hat functions*. The core of the FE method is a Galerkin procedure: the transformation of a system of PDEs into one of ODEs via a projection onto the set of shape (or hat) functions. The decomposition ansatz used in FE methods can be compared to other decompositions like the ones used as *spectral methods* for solving problems in quantum mechanics, elasticity, or vibration. In spectral methods the test functions are often Fourier series, Legendre polynomials, or the like, which means they are oscillatory functions being nonzero along the whole domain Ω, and they form an orthonormal basis, i. e. $\int v_i v_j \,\mathrm{d}\Omega$ equals 1 only if $i=j$ and else 0. The orthogonality of the basis functions is the useful property turning all but one component of integrals of the form $\int v_i \sum_j a_j v_j \,\mathrm{d}\Omega$ to zero. Note that if one was using hat functions as the v_i one would get a similar effect: integrals of the form $\int v_i v_j \,\mathrm{d}\Omega$ would generally evaluate to zero unless $|j-i| \leq 1$ and the two hats overlap. In spectral methods the complexity and fineness of the structure that can be modelled depends on the number of test functions and their maximum spatial frequency. In the FE method the finest modelling scale depends on both, the grid size and the order of shape functions used. One can see that describing a vibrating guitar string either in terms of sine and cosine basis functions or in a finite element framework are two approaches which are mathematically closely related but are quite different in how the approaches mirror the physics.

Here, $\{N'\}$ are the shape functions for the nodal displacement u_e, $\{n\}$ stands for a surface normal on $\partial\Gamma$, and the matrix $[B]$ is the shorthand for the product $\{L\}\{N\}^T$ with $\{L\}^T = (\partial_x, \partial_y, \partial_z)$. The upper index P added here to the stiffness and mass matrices makes sense in a complex FE system where it designates that the corresponding matrix encompasses the subsystem of finite elements for solving the wave equation with the DOF P. The lower index e designates an elemental equation system. Under a close look it can be seen that equation K.5 is analogue equation K.4, only that the term δP, after having first replaced each v, has later been factored out. (Some intermediate steps are given in [8].) The other aspect that can be seen is how the last term on the LHS of equation K.4 is used to connect the oscillating surface of a structural part with the sound field in the acoustic domain next to it in the FE model (equation K.5). Looking at the second derivative of the displacement it can be imagined that the acceleration of a structural surface pushes or pulls on the medium in the acoustic domain and leads to a pressure rise or lowering. This is how ANSYS handles one direction of fluid-structure interaction (FSI) in acoustic problems, the other direction is outlined in section K.3.

It is instructive to compare this abstractly written end result in the following with several different FE simulation types because it shows how this type of shorthand for the assembled ODE system can be read almost like an equation of motion for the simulation. The naming convention of calling $[M_e^P]$ the mass matrix and $[K_e^P]$ the stiffness matrix follows that line. The row of examples can be started by going back from a stationary harmonic analysis to a transient system, i. e. return from the Helmholtz to the wave equation, resulting in

$$[M_e^P]\{\ddot{P}_e\} + [K_e^P]\{P_e\} + \varrho_0[R_e]\{\ddot{u}_e\} = \{0\}. \tag{K.6}$$

Adding damping in the form of sound dissipation at the domain boundary will make the system look like

$$[M_e^P]\{\ddot{P}_e\} + [C_e^P]\{\dot{P}_e\} + [K_e^P]\{P_e\} + \varrho_0[R_e]\{\ddot{u}_e\} = \{0\} \tag{K.7}$$

with

$$[C_e^P] = \frac{\beta}{c}\int_\Gamma \{N\}\{N\}^T \,\mathrm{d}\Gamma \tag{K.8}$$

where the constant $\beta = \frac{r}{\varrho_0 c}$ receives the function of a boundary absorption coefficient, whereby r describes absorption at the boundary [8].

K.2 FE formulation of the structure

The abstract matrix notation of the ODE system representing the FE model used for a transient structural analysis can be written as

$$[M]\{\ddot{u}\} + [C]\{\dot{u}\} + [K]\{u\} = \{F^a\} \tag{K.9}$$

where $\{u\}$ is the nodal displacement vector and $\{F^a\}$ contains the applied nodal load forces. $[M]$, $[C]$, and $[K]$ are the mass, damping, and stiffness matrices. For a

harmonic response analysis an ansatz $\{u\} = (\{u_1\} + i\{u_2\})e^{i\omega t}$ can be made for the displacement and an analogue one for the force, which turns the ODE system into

$$\left(-\omega^2[M] + i\omega[C] + [K]\right)\ (\{u_1\} + i\{u_2\}) = (\{F_1\} + i\{F_2\}). \tag{K.10}$$

Equations K.9 and K.10 seem logic extensions of the shorthand language presented in the equations above. But in fact they are the very abstractly written end results of independently deducing an FE model of elastic solids.

At the basis of the simulation of structural materials lies the equation of motion of elastic solid matter (the PDE of linear elastodynamics) which can be gained by formulating the local force equilibrium for an infinitesimally small test volume under the assumption of geometrically linear displacements.

$$\varrho\ddot{\boldsymbol{u}} = \operatorname{div}\boldsymbol{\sigma} + \varrho\boldsymbol{b} \tag{K.11}$$

$$\boldsymbol{\sigma} = \boldsymbol{\mathcal{C}}\boldsymbol{\varepsilon} \tag{K.12}$$

$$\boldsymbol{\varepsilon} = \nabla^{\mathrm{sym}}\boldsymbol{u} = \frac{1}{2}(\nabla\boldsymbol{u} + \nabla^T\boldsymbol{u}) \tag{K.13}$$

Here, the term $\varrho\boldsymbol{b}$, containing the vector field $\boldsymbol{b}$, represents the load forces per volume. $\operatorname{div}\boldsymbol{\sigma}$ is the term responsible for accelerating an infinitesimal volume if the stress $\boldsymbol{\sigma}$, measured in the units of force per area, is not pulling with equal strength from two opposing sides. The stress state $\boldsymbol{\sigma}$ is put in relation to the strain field $\boldsymbol{\varepsilon}$ via the elasticity tensor (or stiffness tensor) $\boldsymbol{\mathcal{C}}$. Equation K.12 is called material law or constitutive law. The strain tensor $\boldsymbol{\varepsilon}$ reflects the deformation state and is composed of displacement differentials, the particular expression above contains the assumption of geometrical linearity. (Appendix M lists the basic definitions and writing conventions for these tensors and outlines how the above tensor equations can be written as equations involving 2D matrices.)

The authors of [8] however, do not base the deduction of the FE equation system K.9 on the force or momentum balance to be rewritten in the weak form. Instead, they start out from a local energy balance which is formulated in the framework of virtual variations δ and thus already written in the weak form (because the list of nodal virtual displacements fulfills the purpose of the test function under the integral). The initial statement is the principle of virtual work

$$\delta U = \delta V \tag{K.14}$$

and means that any virtual change in the internal strain energy U stored in a test volume of solid matter must come through work V done to the test volume by external mechanisms. For the general case they split U up into two contributions U_1 and U_2 where the first stems from a volume integral over $\delta\varepsilon\,\sigma = \delta\varepsilon\,C\varepsilon$ and the second one from a surface integral accounting for surfaces moving against a so-called foundation stiffness acting as distributed resistance [8]. The latter is a computational tool used in beam and shell elements, so it is not of relevance for the present discussion. Thus δU evaluates to

$$\delta U = \delta U_1 = \int_\Omega \delta\varepsilon\, C\varepsilon\, \mathrm{d}\Omega = \int_{\Omega_e} \{\delta\varepsilon\}[C]\{\varepsilon\}\, \mathrm{d}\Omega_e = \{\delta u\}^T \int_{\Omega_e} [B]^T[C][B]\, \mathrm{d}\Omega_e\, \{u\}$$

Hereby, the transition from ε to $\{\varepsilon\}$ is the one from continuum mechanics to the discretised finite-element framework, so the strain can be expressed in terms of nodal displacements $\{\varepsilon\} = [B]\{u\}$ with the help of the strain-displacement matrix $[B]$ which is based on the element shape functions. The virtual work δV, on the other hand, consists of three contributions, the work against the forces of inertial acceleration (also called d'Alembert forces in the context of variational calculus) δV_1, the work done by pressure forces δV_2, and by externally applied loads δV_3.

$$\begin{aligned}
\delta V_1 &= -\int_\Omega \delta u\, \varrho \ddot{u}\, \mathrm{d}\Omega = -\int_{\Omega_e} \{\delta w\}^T \varrho \{\ddot{w}\}\, \mathrm{d}\Omega_e = -\{\delta u\}^T \varrho \int_{\Omega_e} [N']^T [N']\, \mathrm{d}\Omega_e\, \ddot{u} \\
\delta V_2 &= \int_\Gamma \delta u \cdot \hat{n} p\, \mathrm{d}\Gamma = \int_{\Gamma_e} \{\delta w_n\}^T \{P\}\, \mathrm{d}\Gamma_e = \{\delta u\}^T \int_{\Gamma_e} [N'_n]\{P\}\, \mathrm{d}\Gamma_e \\
\delta V_3 &= \int_\Omega \delta u\, \varrho b\, \mathrm{d}\Omega = \{\delta u\}^T \{F_e^{nd}\}^T
\end{aligned}$$

The rule that has been used here for the transition into the FE framework is the relation between the displacements within the element $\{w\}$ and the nodal displacements[2] $\{w\} = [N']\{u\}$ via the matrix of shape functions. $[N'_n]$ is the matrix of shape functions for motions normal to the surface. Writing all out in the finite element framework and eliminating the arbitrary variations in u yields

$$[K_e]\{u\} = -[M_e]\{\ddot{u}\} + \{F_e^{pr}\} + \{F_e^{nd}\} \tag{K.15}$$

with

$$\begin{aligned}
[K_e] &= \int_{\Omega_e} [B]^T [C][B]\, \mathrm{d}\Omega_e, \\
[M_e] &= \varrho \int_{\Omega_e} [N']^T [N'][B]\, \mathrm{d}\Omega_e, \\
\{F_e^{pr}\} &= \int_{\Gamma_e} [N'_n]^T \{P\}\, \mathrm{d}\Gamma_e, \text{ and} \\
\{F_e^{nd}\} &= \text{ the nodal forces applied to the element.}
\end{aligned}$$

Compared with the FE system written above in equation K.9 just one detail seems to be missing. In order to get that equation, a matrix of velocity-proportional damping coefficients must still be added. That step is often lacking any microscopical motivation and is just the ad hoc emulation of the well-known concept of a damped oscillator. The damping constants are then supplied by empirical measurements, where the procedure of postprocessing the experimental data reflects the ad hoc assumptions.

K.3 Fluid-structure-interaction (FSI)

The last term on the LHS in equation K.5 describes how a displacement of the boundary of the acoustic domain is connected with the pressure DOF inside the

[2] In the ANSYS documentation [8] always the set of shape functions from the currently discussed domain are the set labelled $[N]$ and the shape functions of the foreign domain are $[N']$. Here, in order to stay consistent, the pressure shape functions are labelled $[N]$ and the displacement ones are $[N']$.

acoustic domain. With this one can describe a church bell where the metal's vibration leads to emanating pressure waves in the air. The other direction, e. g. how a pressure wave carried through the air sets our eardrums in motion is the term $\{F_e^{pr}\}$ in equation K.15 which can also be written as

$$\{F_e^{pr}\} = \int_{\Gamma_e} \{N'\}^T \{P\} \{n\} \, \mathrm{d}\Gamma_e = \int_{\Gamma_e} \{N'\} \{N\}^T \{n\} \, \mathrm{d}\Gamma_e \, \{P_e\} = [R_e] \{P_e\}$$

which means that an additional surface force term can be superimposed acting on faces of structural elements in contact with an acoustic domain and that this normal force is caused by the sound pressure.

K.4 Piezoelectricity

A salt crystal consists of a regular grid of sodium and chloride atoms. It is the symmetry of this atomic crystal lattice and the symmetry of the electron orbitals around its atoms which ensure that there is no net shift of the grid of valence electrons against the grid of ions if a sample volume of salt is subjected to a small (and geometrically linear) elastic deformation. But for other crystal lattices these symmetries are broken, e. g. in the case of quartz (SiO_2). The consequence of an average shift of the valence electron orbital structure against the ion lattice is that under elastic deformation these crystals exhibit a varying polarisation and one can measure strong charge accumulations on the surfaces. The inverse effect is that an external electric field tearing the electron and ion systems into different directions is able to induce internal stresses and deform a piezoelectric crystal. Since the discovery of the piezoelectric effect in tourmaline crystals in 1880 by Jacques and Pierre Curie, materials exhibiting a stronger piezoelectric effect have been searched and developed. The standard piezoelectric components used for technical applications nowadays are polycrystalline ceramics (e. g. perovskites like barium titanate ($BaTiO_3$), lead titanate ($PbTiO_3$), or lead zirconate titanate ($PbZr_xTi_{1-x}O_3$)). There are many applications of piezoelectric components from everyday devices like a lighter where a hammer hits a small piezo crystal to produce the spark igniting the flame to microscopic positioning systems used in research where voltages are used to control component deflections. An important field of application for over a century has been sonar systems where piezo components are being used as both sound emitters and receivers.

Formal description

In an elastic solid without piezoelectric effect the local stress state is a function of the strain (see appendix M for a more detailed explanation of the symbols and equations.)

$$\boldsymbol{\sigma} = \boldsymbol{\mathcal{C}} \boldsymbol{\varepsilon}. \tag{K.16}$$

But in a piezoelectric material the stress depends not only on the strain because there is the additional contribution caused by the electric field $\boldsymbol{E}$ and conveyed by the tensor $\boldsymbol{e}$, the *piezoelectric stress tensor*:

$$\boldsymbol{\sigma} = \boldsymbol{\mathcal{C}} \boldsymbol{\varepsilon} + \boldsymbol{e} \boldsymbol{E}. \tag{K.17}$$

The inverse effect is described by

$$\boldsymbol{D} = \boldsymbol{e}^T\boldsymbol{\varepsilon} + \boldsymbol{\epsilon}\boldsymbol{E} \tag{K.18}$$

which has to be compared with

$$\boldsymbol{D} = \boldsymbol{\epsilon}\boldsymbol{E} = \epsilon_0\boldsymbol{E} + \boldsymbol{P} = \epsilon_0(1+\boldsymbol{\chi})\boldsymbol{E} = \epsilon_0\boldsymbol{\epsilon}_r\boldsymbol{E} \tag{K.19}$$

for a material without piezoelectric effect. It means in a piezoelectric material the electric displacement field $\boldsymbol{D}$ does not only depend on the $\boldsymbol{E}$-field but also on the deformation state.

The constant $\epsilon_0 = 8.854\,187\,817 \times 10^{-12}\,\mathrm{F/m}$ is the *vacuum permittivity* or *electric constant*. $\boldsymbol{P}$ is the polarisation vector which gives the density of dipole moment and is a measure of the spatial offset between the lattices of negative and positive charge carriers. $\boldsymbol{\epsilon}_r = (1+\boldsymbol{\chi})$ is the relative permittivity, which accounts for the contribution of the polarisation to $\boldsymbol{D}$. $\boldsymbol{\chi}$ is called the electric susceptibility. In an isotropic dielectric medium $\boldsymbol{\epsilon}_r$ it is just a scalar (like ϵ_0) and a measure for how much the induced polarisation increases $\boldsymbol{D}$ inside the dielectric matter. But if the spring constants for offsetting negative and positive charges against each other along the different dimensions are not equal, then $\boldsymbol{\epsilon}_r$ turns into a tensor to allow directions of $\boldsymbol{P}$ which are not parallel to $\boldsymbol{E}$. Such behaviour is also the case in piezoelectric materials. Analoguely, the constitutive tensor $\boldsymbol{\mathcal{C}}$ of a (not isotropic) piezoelectric solid induces tension ellipsoids not aligned with the strain states. And finally, the piezoelectric stress tensor $\boldsymbol{e}$ is there to translate between nonaligned electric fields and stress states.

The behaviour of a piezoelectric material in the physically linear range can be fully characterised by giving the matrices of $\boldsymbol{\mathcal{C}}$, $\boldsymbol{e}$, and $\boldsymbol{\epsilon}_r$. For all these tensor equations a shorthand based on 2D matrices exists which is explained in appendix M. In that notation, and if the polarisation of the piezoelectric material is parallel to the z-axis, one has

$$\boldsymbol{C} = \begin{pmatrix} c_{11} & c_{12} & c_{13} & 0 & 0 & 0 \\ c_{12} & c_{11} & c_{13} & 0 & 0 & 0 \\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{66} \end{pmatrix} \quad \boldsymbol{\epsilon}_r = \begin{pmatrix} \epsilon_1 & 0 & 0 \\ 0 & \epsilon_1 & 0 \\ 0 & 0 & \epsilon_3 \end{pmatrix} \quad \boldsymbol{e} = \begin{pmatrix} 0 & 0 & e_{31} \\ 0 & 0 & e_{31} \\ 0 & 0 & e_{33} \\ 0 & e_{15} & 0 \\ e_{15} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

In equation K.17 & K.18 one calculates the stress and the $\boldsymbol{D}$-field induced by strain and $\boldsymbol{E}$-field. Another form of the equation system often needed is the pair

$$\boldsymbol{\varepsilon} = \boldsymbol{S}\boldsymbol{\sigma} + \boldsymbol{d}\boldsymbol{E} \tag{K.20}$$

$$\boldsymbol{D} = \boldsymbol{d}^T\boldsymbol{\sigma} + \boldsymbol{\epsilon}\boldsymbol{E} \tag{K.21}$$

where $\boldsymbol{S} = \boldsymbol{C}^{-1}$ is the elastic compliance and $\boldsymbol{d}$ the *piezoelectric strain tensor*. Here one calculates the strain and $\boldsymbol{D}$-field caused by a given setting of stress and $\boldsymbol{E}$-field. Furthermore, it makes a difference whether $\boldsymbol{C}$ and $\boldsymbol{S}$ are measured at constant $\boldsymbol{E}$ or constant $\boldsymbol{D}$, and similarly, whether $\boldsymbol{\epsilon}_r$ is measured at constant stress or strain. The quantity kept constant is usually indicated as a superscript, e. g. $\boldsymbol{C}^E$ or $\boldsymbol{\epsilon}_r^{\boldsymbol{\sigma}}$.

K.5 Damping

Damping is a topic where there is a large discrepancy between the multitude of microscopic phenomena at its roots and the few simple and popularised models describing macroscopic mechanical systems. Among the simplified mathematical models, the conceptual approaches, and thus, the mathematics differ drastically. In the ANSYS FEM suite, only one of the models, namely viscous damping, is implemented for use with harmonic or fully transient analyses. But some features of other damping models can be emulated by the computation of equivalent viscous damping coefficients. The following section together with appendix N provides the necessary basics.[3]

K.5.1 The viscously damped free harmonic oscillator

A generic harmonic oscillator is described by the differential equation

$$m\ddot{x} + \gamma\dot{x} + kx = f \tag{K.22}$$

where m is the mass, k the spring constant, and $f = f(t)$ a force acting on the mass. The constant γ acts as a damping term. If it is zero, then harmonic oscillations of eternally constant amplitude of the form $x(t) = X\sin(\omega_n t)$ are solutions to the equation once the excitation force is switched to zero, where $\omega_n = \sqrt{k/m}$ is the resonance frequency (also called natural frequency) of the undamped oscillator. The equation can be brought into a simpler form by setting

$$\gamma = 2m\gamma' \qquad \text{and} \qquad k = \omega_n^2 m, \tag{K.23}$$

which turns the equation of motion into

$$\ddot{x} + 2\gamma'\dot{x} + \omega_n^2 x = 0 \tag{K.24}$$

where f has been set to zero for the purpose of looking at the form of motions when the external influence has been stopped. Now, the ansatz $x(t) = e^{\lambda t}$ will lead to a quadratic equation in λ

$$\lambda^2 + 2\gamma'\lambda + \omega_n^2 = 0 \qquad \text{with the solutions} \qquad \lambda_{1,2} = 2\gamma'^2 \pm \sqrt{\gamma'^2 - \omega_n^2}.$$

This means that equation K.24 has solutions of the form

$$x(t) = X_1 e^{\lambda_1 t} + X_2 e^{\lambda_2 t} = e^{2\gamma'^2 t}\left(X_1 e^{\sqrt{\gamma'^2-\omega_n^2}t} + X_2 e^{-\sqrt{\gamma'^2-\omega_n^2}t}\right).$$

The constants X_1 and X_2 can be freely chosen to match any choice of initial conditions $x(t=0)$ and $\dot{x}(t=0)$. If $\gamma'^2 - \omega_n^2 < 0$, then the terms in parentheses have purely imaginary exponents and yield harmonic oscillations. If $\gamma' > 0$ then the oscillation amplitude decays exponentially due to the factor $e^{2\gamma'^2 t}$, this regime is called

[3] The book 'The mechanics of vibration' by Bishop and Johnson [46] has been found to offer very useful explanations about the basics of damped oscillating mechanical systems. Therefore, the nomenclature and the definition of some important terms and concepts have been adapted from it.

underdamped. If $\gamma'^2 - \omega_n^2 > 0$, then all three exponents are real-valued, the regime is *overdamped*, and after excitation the oscillator shows no oscillation, but only a slowing down and moving to the equilibrium position. The amount of damping at the boundary between the two regimes, where an overshooting of the equilibrium sets in (or vanishes), is called *critical damping* and the corresponding value of γ is labelled γ_{cr}. The ratio of the actual viscous damping constant to the critical value

$$\zeta = \frac{\gamma}{\gamma_{\mathrm{cr}}}$$

is called the *damping ratio* (or *damping factor*) [446].

K.5.2 The viscously damped forced harmonic oscillator – stationary solution

The section above has the intention to recall, that it is the velocity-proportional friction term in the equation of motion K.22 that is responsible for damping. What are the consequences of such a damping term in a forced harmonic analysis where the stationary response to a harmonic forcing is computed? In order to find that out for the harmonic oscillator, one has to apply a force $f(t) = Fe^{i\omega t}$ in eq. K.22 and make the ansatz

$$x(t) = Xe^{i\omega t + \varphi}. \tag{K.25}$$

In the resulting equation

$$-m\omega^2 X + i\omega\gamma X + kX = Fe^{-i\varphi} = F(\cos\varphi - i\sin\varphi) \tag{K.26}$$

one can compare the real and imaginary parts

$$\frac{X}{F}(k - m\omega^2) = \cos\varphi \qquad \text{and} \qquad -\frac{X}{F}\omega\gamma = \sin\varphi.$$

Making use of the identities $\cos^2\varphi + \sin^2\varphi = 1$ and $\tan\varphi = \frac{\sin\varphi}{\cos\varphi}$ one can finally get to the expressions for the characteristic dependencies of amplitude X and phase φ on the frequency and how they depend on the damping parameter γ.

$$X(\gamma,\omega) = \frac{F}{\sqrt{(k - m\omega^2)^2 + \omega^2\gamma^2}} \tag{K.27}$$

$$\varphi(\gamma,\omega) = \arctan\left(\frac{\omega\gamma}{m\omega^2 - k}\right) \tag{K.28}$$

By replacing m with k/ω_n^2 these can be rewritten as

$$X(\gamma,\omega) = \frac{F}{\sqrt{k^2(1 - (\frac{\omega}{\omega_n})^2)^2 + \omega^2\gamma^2}} \tag{K.29}$$

$$\varphi(\gamma,\omega) = \arctan\left(\frac{\omega\gamma/k}{(\frac{\omega}{\omega_n})^2 - 1}\right). \tag{K.30}$$

Figure K.3 shows a plot of this characteristic frequency response.

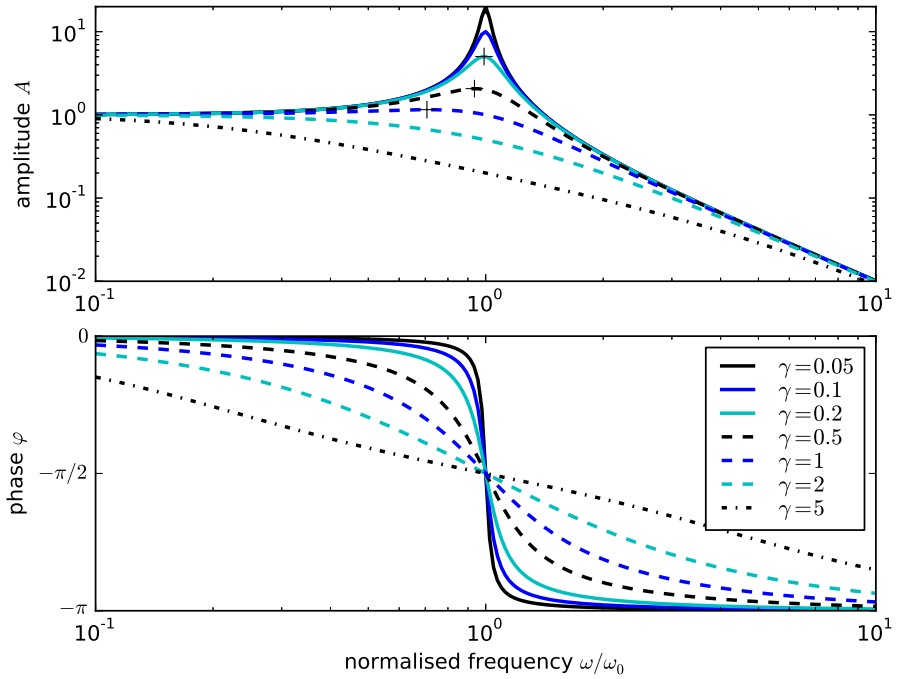


Figure K.3 The frequency response of a harmonic oscillator with viscous damping. In the above plot the frequency axis has been normalised to the undamped resonance frequency $\omega_n = \sqrt{k/m}$. The parameters F and k have been set to 1. It can be seen that viscous damping reduces the resonance frequency. To connect the abstract plot with experience, one can think of the example of a car bumping along an uneven road with ripples and cracks, but generally flat on the wider scale, where the car follows and amplifies the up and down movements at lowest speeds, but stays level while only the wheels bounce at elevated speeds. A damped harmonic oscillator can be seen or used as a low-pass filter, higher frequencies don't reach the car body. (Analogue plots for the case of ideal hysteretic damping are shown in appendix N.3 as a comparison.)

K.5.3 Measures of damping

Several important quantities can be defined independently of the particular damping mechanism or model and solely based on (a) characterising the temporal decay of free transient oscillatory motion, (b) relative amplitudes in the case of the stationary response to harmonic forcing or (c) fundamental energy considerations. (See also e. g. [446].) First, there is the *loss factor* η which is the ratio between the energy dissipated per oscillation cycle (2π) and the total mechanical energy still stored in the system at that moment, i. e.

$$\eta = \frac{\Delta W}{2\pi E_{\text{tot}}} = \frac{\Delta W}{2\pi(E_{\text{kin}} + E_{\text{pot}})} = \frac{\Delta W}{2\pi(T+U)} = \frac{\Delta W}{2\pi U_{\max}} = \frac{\Psi}{2\pi}. \tag{K.31}$$

The variable Ψ stands for an alternative measure called *specific damping capacity*. It is the relative amount of energy dissipated during one oscillation cycle:

$$\Psi = \frac{\Delta W}{W} = \frac{\Delta W}{T+U} = \frac{\Delta W}{U_{\max}}. \tag{K.32}$$

The damping can also be quantified by the ratio of amplitudes $\frac{X_{n+1}}{X_n}$ of two consecutive oscillation amplitudes. Thus the *logarithmic decrement of free vibrations* is defined by [39]

$$\delta_{\mathrm{ld}} = \ln \frac{X_n}{X_{n+1}}. \tag{K.33}$$

(The lower index "ld", short for logarithmic decrement, is added here to the usual symbol δ in order to distinguish it from loss angles labelled with the symbol δ, or δ_{loss} for better distinction.) For experimental determination the formula $\delta_{\mathrm{ld}} = \frac{1}{m} \ln \frac{X_n}{X_{n+m}}$ can be useful. Finally, the *Q-factor* of a resonating system (the *quality factor* or *resonant amplification factor* [446]) can be taken from measurements on stationary oscillations via the formula

$$Q = \frac{\omega_n}{\omega_2 - \omega_1} \tag{K.34}$$

where ω_n is the frequency of a resonance peak in the amplitude response of a suitable physical quantity and ω_1, ω_2 are the points downwards and upwards on the frequency axis where that amplitude has decreased by a factor $1/\sqrt{2}$ from the peak value. This definition is applicable when the frequencies ω_1 and ω_2 correspond to working points where the dissipation power of the damping mechanism is half its power at the resonance peak ω_n. That is the case for many damping models when ΔW is proportional to the square of the displacement or strain amplitude and when by consequence the reduction of the amplitude by a factor of $1/\sqrt{2}$ cuts ΔW in half. In these cases any physical quantity with an amplitude proportional to the displacement amplitude is suitable for inferring Q, e. g. voltage, current, charge, velocity, pressure in many transducers and resonators. Since $20 \log_{10}(\frac{1}{\sqrt{2}}) = -3.01\,\mathrm{dB}$, the denominator of equation K.34 is called the *3 dB bandwidth* [321].

Sun & Lu [446] list the following conversions valid in the regime of weak damping where $\zeta < 1$, $\eta < 1$, $\delta_{\mathrm{ld}} < 1$, $2Q > 1$, $\Psi < 4\pi$:

$$\begin{aligned}
\delta_{\mathrm{ld}} &= 2\pi\zeta = \frac{\pi}{Q} = \pi\eta \\
Q &= \frac{\omega_n}{\omega_2 - \omega_1} = \frac{1}{2\zeta} = \frac{1}{\eta} \\
\eta &= 2\zeta = \frac{\delta_{\mathrm{ld}}}{\pi} \\
\Psi &= \frac{2\pi}{Q} = 2\pi\eta = 4\pi\zeta = 2(\delta_{\mathrm{ld}} - \delta_{\mathrm{ld}}^2), \text{ or } 2\delta_{\mathrm{ld}} \\
Q &\approx \frac{\pi}{\delta_{\mathrm{ld}} - \delta_{\mathrm{ld}}^2}
\end{aligned} \tag{K.35}$$

K.5.4 Frequency-independent damping

In the viscous damping model the friction force is proportional to the velocity. Consequently, for the particular case of harmonic vibration the energy lost per oscillation cycle

$$\Delta W = \oint \gamma \dot{x}\, \mathrm{d}x = \gamma \oint \dot{x}^2\, \mathrm{d}t = \gamma \int_0^{\frac{2\pi}{\omega}} \omega^2 X^2 \cos^2 \omega t\, \mathrm{d}t = \pi\gamma|\omega|X^2 \tag{K.36}$$

is proportional to the frequency. Oscillating systems with this property can indeed be well reproduced in a lab setup, e. g. with a sidearm of a pendulum submerged in oil so that the force slowing the pendulum down is created by laminar flow phenomena and thus proportional to the velocity. But in the case of vibrating structures made of solid materials where the damping comes from imperfect elasticity, the experimental data looks different. It has been shown (an early work on glass, wood, rubber, and several metals can be found in [233]) that ΔW is indeed proportional to the square strain amplitude, yet almost independent of ω for many common construction materials.

The viscous damping model can a posteriori be made to reflect that result by assuming $\gamma \propto 1/\omega$. So replacing $\gamma\dot{x}$ with $h\dot{x}/\omega$ in equation K.36 yields

$$\Delta W = \pi h X^2. \tag{K.37}$$

For treating the case of stationary forced harmonic oscillation (by utilising the ansatz $f(t) = Fe^{i\omega t}$ and $x(t) = Xe^{i(\omega t+\varphi)}$) it means that instead of

$$m\ddot{x} + \gamma\dot{x} + kx = f \tag{K.38}$$

turning into

$$\left(-m\omega^2 + i\omega\gamma + k\right) X = Fe^{-i\varphi} \tag{K.39}$$

one has

$$m\ddot{x} + h\dot{x}/\omega + kx = f \tag{K.40}$$

turning into

$$\left(-m\omega^2 + (k + ih)\right) X = Fe^{-i\varphi}. \tag{K.41}$$

It can be seen, that equation K.41 might as well have come from an analogue treatment of the differential equation

$$m\ddot{x} + kx = f \qquad \text{with } k = k' + ih = k' + ik'' = k'(1 + i\eta), \tag{K.42}$$

where a complex-valued stiffness $k = k' + ik''$ has been defined. This approach of employing a complex-valued stiffness is often referred to as *structural* or *hysteretic* damping, and it is popular besides the above-stated reasons for the fact that it is a linear differential equation just like equation K.22/K.38. In the frequency domain a similar treatment of equation K.41 as conducted above for the viscous case leads to the response functions plotted in figure K.4. The derivation can be found in appendix N.3.

The only problems with equation K.42 arise in the time domain where it describes an unphysical system [216], as in this case the force law $f(t) = kx(t)$ would imply that a real-valued displacement trajectory leads to a complex-valued force. Another facet is that the computation of time-domain response functions of equation K.42 to impulse excitations by means of Fourier series yields so-called *impulse response precursors, IRP*, i. e. the system begins to react even before the excitation happens, which violates causality [92]. In order to be able to describe transient damped motions under the above-stated requirements that the damping be independent of deformation rate and proportional to the square deformation amplitude,

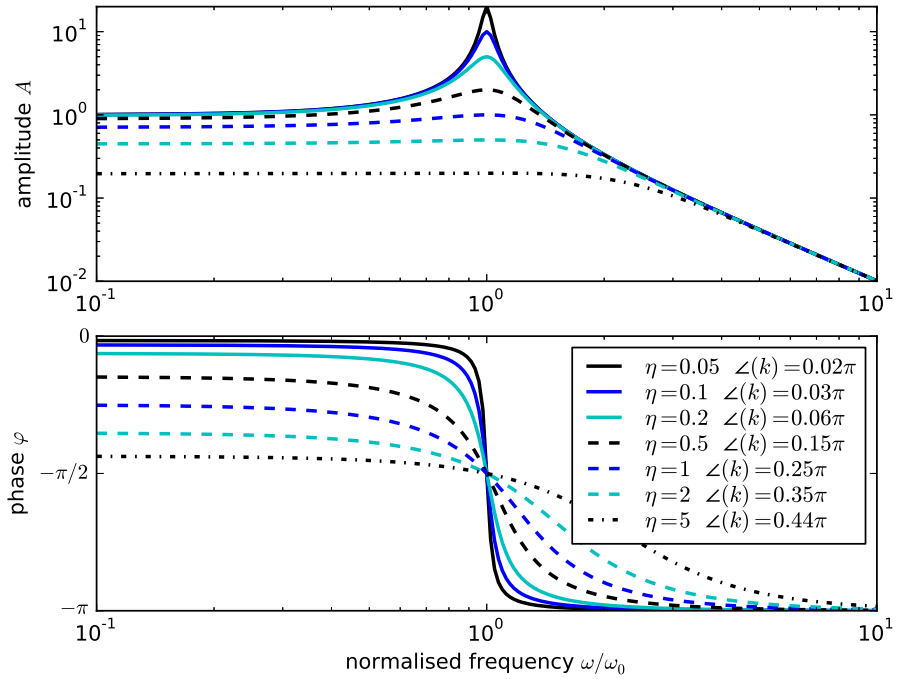


Figure K.4 The frequency response of a harmonic oscillator with ideal hysteretic damping.

In the above plot the frequency axis has been normalised to the undamped resonance frequency $\omega_0 = \sqrt{k'/m}$. The parameters F_a and k' have been set to 1. The upper plot shows the frequency response of the amplitude, the lower plot the phase response. The varied parameter is the loss factor η which in the case of hysteretic damping can be identified with k''/k', the ratio between imaginary and real parts of the complex stiffness. It can be seen that the resonance frequency (peak amplitude) does not depend on η. From equation N.21 it can be taken that for $\omega/\omega_0 \to 0$ the phase offset φ approaches $\arctan\eta$ which means it directly mirrors the loss angle δ_{loss} in that regime.

different versions of the equation of motion must be sought. (Examples are shown in appendix N.) Such differential equations will not necessarily have sinusoidal displacement trajectories among their solutions. Therefore it is stressed that, firstly, as long as the concept of vibration damping by a complex stiffness originates from nothing else but rewriting the viscous damping term as $-h\dot{x}/\omega$ it is exclusively defined for use in the frequency domain, and secondly, that for increasing levels of damping the assumption of sinusoidal variations of physical quantities might become less and less realistic.

In a force-position diagram (as in figure N.1) cyclic trajectories which are collapsed to a line correspond to the absence of energy dissipation, and correspondingly, any kind of energy dissipation leads to hysteresis loops encircling an area of size equivalent to ΔW. Therefore, the naming convention of calling a damping model sufficing equation K.37 *hysteretic damping* is somewhat suboptimal. Other names, several of them not much less suboptimal, are [315] *structural*, *material*, *rate-independent*, *ideal hysteretic* [92], or *viscoelastic* [446] damping.

K.5.5 The loss factor and complex moduli

In the framework of material descriptions with complex moduli the loss factor η can be identified with the imaginary-to-real moduli ratio according to the equation

$$\eta(\omega) = \frac{\Delta W(\omega)}{2\pi U_{\max}} = \frac{k''(\omega)}{k'(\omega)} = \frac{E''(\omega)}{E'(\omega)} = \frac{1}{Q(\omega)} = \tan\delta_{\text{loss}}(\omega). \tag{K.43}$$

Deductions of that equation can be found in [321] (p. 47ff) and [158]. Furthermore, equation K.43 says that the angle found in the complex plane between k (or E) and the abscissa can be identified with the phase lag of a system's response $x(t) = X\sin(\omega t + \delta_{\text{loss}})$ to harmonic excitation $f(t) = F\sin\omega t$. In such a case the *loss angle* $\delta_{\text{loss}}(\omega)$ is equivalent to the loss factor $\eta(\omega)$ in its information content and can alternatively be given to characterise the damping behaviour of the system.

K.5.6 Many-DOF systems and Rayleigh damping

A system representing a linear elasticity problem with many degrees of freedom can for example be a chain of masses connected with springs moving on a single rail, a few masses connected with rubber bands and bouncing in three dimensions, or a truss construction. If there are only linear springs involved then the equation of motion can be written as

$$[m]\{\ddot{x}\} + [k]\{x\} = \{f\}. \tag{K.44}$$

This might as well be a FE system. All the physics of the system is expressed within the two matrices $[m]$ and $[k]$. How can damping be introduced into such a system? In the case of viscous damping the two most obvious options might be adding either a term $\gamma\{\dot{x}\}$ with the same damping constant for all point masses or a term $[\gamma]\{\dot{x}\}$ with individual constants. The drawback of oversimplification stands against the drawback of having to know or determine many damping constants. Rayleigh damping means the setting

$$[\gamma] = \alpha[m] + \beta[k] \qquad \text{with } \alpha, \beta \in \mathbb{R}^+ \tag{K.45}$$

and allows to implement many different distributions of damping strengths within a network of springs by tuning the weighting of two constants α and β. Of course the one distribution closest to reality might not be in that set. In fact, the motivation of Rayleigh damping lies not in any physical background. It comes purely from advantages of mathematical nature of the decomposition of equation K.45 which are useful particularly in connection with modal analyses, and which are outlined in appendix N.

Now that one single damping constant for the system has been split up into many, all the relations involving η or ζ given in equations K.35 are of course not applicable any more. The question is, whether they can be replaced with relations based on α and β instead. Indeed it turns out that the coefficient α has a stronger effect on lower frequencies and β has a damping power increasing with frequency. This frequency-dependence is also a result of the type of eigenmode analysis becoming

possible through the decomposition of equation K.45, so it is given as the damping ratio of the i^{th} eigenmode

$$\zeta_i = \frac{\alpha}{2\omega_i} + \frac{\beta\omega_i}{2} \tag{K.46}$$

where ω_i is the eigenfrequency of the mode. In practice this formula is also commonly applied outside the context of a modal analysis to set the damping ratio of a simulated model to a desired value at the frequency or in the frequency band of interest [8]. Figure K.5 shows a plot of the frequency-dependence of α- and β-damping and the sum given by equation K.46. It also illustrates how to choose a suitable setting of the two weights in order to minimise the damping level variation within a frequency interval of interest.

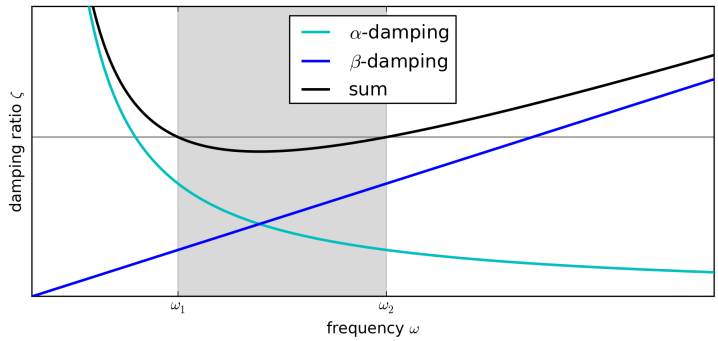


Figure K.5 Rayleigh damping components.
Formula K.46 gives the relation between the damping ratio and the damping matrix weights α and β that holds for all eigenmodes with their eigenfrequencies ω_i of one multi-DOF system if Rayleigh-damping is implemented. In practice, this formula can also be used in the other direction: by choosing a certain combination of values for the two weights the damping ratio of a simulated model can be determined. If one is interested in simulating only a small frequency band, the simplest approach is to leave one of the two weights at zero and use the other one for the tuning. As the constants α and β have no physical meaning and as the term with β has at least the similarity to ideal hysteretic damping of being stiffness-proportional, the choice often falls on $\beta > 0$ and $\alpha = 0$. In case a wider frequency interval has to be simulated and the variation of the damping ratio should be kept low within that interval, then equation K.46 can be used to establish a situation as for the grey shaded frequency band shown in the diagram, where the damping ratio at the band edges is equal and the variation within is low.

K.5.7 Composition of the damping matrix in ANSYS

Damped FE equation systems are implemented in ANSYS always by an equation system in the form

$$[M]\{\ddot{u}\} + [C]\{\dot{u}\} + [K]\{u\} = \{F\} \tag{K.47}$$

where $\{u\}$ is the list of nodal displacements, $\{F\}$ is the list of load forces, and the matrices $[M]$, $[K]$, and $[C]$ are the mass, stiffness, and damping matrix. The damping matrix $[C]$ is composed, according to the concept of Rayleigh damping, as the sum of terms building on $[M]$ and $[K]$. This sum is given by equation 15-20 in chapter 15.3 of the ANSYS documentation [8] and has the form

$$[C] = \alpha[M] + (\beta + \frac{2}{\omega}\zeta)[K] + \sum_{j=1}^{N_m}\left((\beta_j + \frac{2}{\omega}\zeta_j)[K_j]\right) + P. \tag{K.48}$$

Here, P stands for particular contributions only applicable with specialised elements and analysis types which are not of interest and not discussed here. The sum over $j = 1, 2, ..., N_m$ goes over the various structural materials employed in the model, that means it is a sum over parts $[K_j]$ of the whole stiffness matrix $[K]$ corresponding to the equation subsystem belonging to one material type. The intended use case is that only few summands are switched to be nonzero, for example only α and β to get to the situation of formula K.45. Using the coefficients β_j instead of β allows for making the β-damping material-specific, and using the terms $\frac{2}{\omega}\zeta$ or $\frac{2}{\omega}\zeta_j$ instead of β or β_j (while $\alpha = 0$) allows for emulating a frequency-independent effect of the viscous damping model.

For dielectric damping the command `mp,lsst` is available in ANSYS allowing to specify a dielectric loss tangent. That this command really acts as may be assumed and turns an initially real-valued dielectric tensor $\boldsymbol{\epsilon}$ by multiplication into

$$\boldsymbol{\epsilon}^{\text{eff}} = \boldsymbol{\epsilon}(1 + \tan \delta_{\text{diel}}) \tag{K.49}$$

is somewhat hidden in the program's theory reference in equations 5-86, 5-119, and 5-120.

K.5.8 Microscopic explanations for sound and vibration damping in solids

Vibrations are damped and sound waves attenuated in solids due to internal friction. But from what does internal friction arise if there is only elastic and no plastic deformation? How should friction for compression waves be imagined? The general answer is that any kind of relaxation process can dissipate mechanical energy and produce entropy and that such processes are maximised if the time scale of the relaxation mechanism corresponds to the motion time period. The most fundamental such damping phenomenon is thermoelastic damping [48]. It occurs even in the purest single crystal because the only requirement is a non-zero thermal expansion coefficient. The inverse effect to thermal expansion is heating upon compression. As compression and dilation are part of all longitudinal sound waves and most vibration motions, local temperature gradients are created. The gradients are equilibrated by irreversible heat flows, and therein consists the dissipation mechanism. One can speak of a relaxation of the temperature field. But thinking of two different elastic moduli, an adiabatic and an isothermal one, one can also think of a relaxation of elasticity. The adiabatic modulus applies for quick compressions and dilations whereas the isothermal one applies for slow changes. Due to the time needed for thermal relaxation, on can imagine that upon a quick bending motion the adiabatic modulus relaxes towards the isothermal one asymptotically. This thought experiment helps to understand why thermoelastic damping is frequency-dependent and why there is always a frequency with maximal damping (when the time scales match best) and less damping for higher or lower frequencies. At low frequencies there is almost no deviation from the isothermal modulus which minimises dissipation, and at high frequencies there is simply no time for much heat transport.

These thoughts can be generalised to many other relaxation processes [330] including the following:

- Grain boundary relaxation occurs when changes of the stress state induce the migration of grain boundaries in polycrystalline materials.

- Point defect relaxation is due to the migration of point defects, similarly there are

- dislocation relaxation,

- Zener relaxation, i. e. the relaxation of the orientation of pairs of dilute atoms in a crystal lattice,

- Snoek relaxation, i. e. small atoms jumping between different types of interatom positions in crystal lattices of larger atoms (e. g. carbon in iron), and

- relaxations associated with phase transformations.

All these processes involve energy barriers to be overcome when atoms change their positions in solids. As a consequence, the availability of activation energy plays a role which makes such relaxation processes temperature-dependent. The listed kinds of material damping are called *anelastic* behaviour, not to be confused[4] with the more general term *viscoelasticity*.

Thermoelastic damping can be described in terms of classical physics. But classical physics is always just a limit case of the broader quantum-mechanical description. The quantum-mechanical facets of vibration damping can be discussed when sound waves are seen as the manifestation of many phonons. In the phonon picture thermoelastic damping is the limit case when both the dominant wavelength and the mean free path of thermal phonons are much smaller than the wavelength of the vibration motion [269]. However, in micro- and nano-structured mechanical resonators this is not always the case. Then, the quantum-mechanical nature of phonons can begin to play a role and damping only understood correctly by describing it in terms of phonon-phonon interaction and phonon scattering at crystal defects. The transitions between the group of phonons forming the vibration mode and the thermal phonon bath throughout the resonator and its surrounding have to be examined.

The unfortunate consequence of the fact that material damping arises from a combination of various microscopic phenomena is the high number of determinants (not only frequency, temperature, vibration mode shape but also material purity, treatment history etc.) and the difficulty to find applicable literature data.

[4] A good definition of these terms can be found in [330]. It is based on the three requirements of ideal elasticity: (1) the strain response to each level of applied stress (or vice versa) has a unique equilibrium value, (2) the equilibrium response is achieved instantaneously, and (3) the response is linear. Anelastic behaviour is encountered when only the second condition does not hold. In the case of linear viscoelasticity, conditions (1) and (2) are not given. In a viscoelastic material, the strain state is also history-dependent.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
$[B]$	strain-displacement matrix
$\boldsymbol{b}$	load force field
$\boldsymbol{\mathcal{C}}$	elasticity tensor (constitutive tensor)
$[C]$	damping matrix
c	speed of sound/light
$\boldsymbol{D}$	electric displacement field
$\boldsymbol{d}$	piezoelectric strain tensor
E	energy; Young's modulus
$\boldsymbol{E}$	electric field
$\boldsymbol{e}$	piezoelectric stress tensor
F, f	force
f	generic function
g	generic function
h	hysteretic damping component, i. e. imaginary part of stiffness
h	shape function for a finite element
i	unit value of imaginary numbers
K	bulk modulus
$[K]$	stiffness matrix
K, k	spring constant, stiffness
$\{L\}$	Nabla operator in list notation
M, m	mass
$[M]$	mass matrix
$\hat{n}$	unit vector in the normal direction
$\boldsymbol{P}$	polarisation field
$P(\vec{x})$	stationary sound pressure field
p	pressure
Q	quality ("pointedness" of a resonance peak)
$[R]$	acoustic fluid boundary matrix
r	absorption at the boundary of the acoustic domain; element coordinate
$\boldsymbol{S}$	compliance tensor
T	kinetic energy
t	time
U	internal enery
$u, \boldsymbol{u}$	displacement
V	work
$\hat{V}$	set of test functions
v	test function
$\{w\}$	vector of displacements of a general point
W	work
X	amplitude (of oscillator motion)
x	distance, spatial coordinate

$\{x\}$	quantity in curled brackets for a nodal list (vector)
$[x]$	quantity in square brackets for a matrix

List of Greek quantity symbols

Symbol	Description
α	mass matrix coefficient (Rayleigh damping)
β	boundary absorption coefficient
β	stiffness matrix coefficient (Rayleigh damping)
Γ	boundary of the region/domain under consideration
γ	damping coefficient/factor
δ	infinitesimal step/variation or virtual operator
δ_{ld}	logarithmic decrement of damping
δ_{loss}	normalised damping factor
$\boldsymbol{\varepsilon}$	strain tensor
$\boldsymbol{\epsilon}, \epsilon_0, \boldsymbol{\epsilon}_r$	permittivity, vacuum permittivity, relative permittivity
ζ	damping ratio
η	loss factor
λ	ansatz parameter
ϱ	density
$\boldsymbol{\sigma}$	stress tensor
ϕ	phase angle
$\boldsymbol{\chi}$	electric susceptibility
Ψ	specific damping capacity
Ω	volume, i. e. spatial region/domain under consideration
ω	angular frequency

List of abbreviations

Abbreviation	Description
DOF	degree of freedom
FE,FEM	finite element (method)
FSI	fluid-structure interaction
IRP	impulse response precursors
LHS	left hand side
ODE	ordinary differential equation
PDE	partial differntial equation
SF	sonofusion

Appendix L

Integration by parts in two and three dimensions (Green's theorem)

In order to be self-contained and because it is not completely trivial, the derivation of the identity used on page 318 shall be shown here. The treatment is based on appendix G of O. C. Zienkiewicz's book on the finite element method [539].

The 1D case: In one dimension the rule of integration by parts states

$$\int_{x_1}^{x_2} uv' \,\mathrm{d}x = [uv]_{x_1}^{x_2} - \int_{x_1}^{x_2} u'v \,\mathrm{d}x$$

and it can be derived from the product rule, since

$$\begin{aligned} \mathrm{d}(uv)' &= v \,\mathrm{d}u + u \,\mathrm{d}v \\ uv &= \int v \;\mathrm{d}u + \int u \;\mathrm{d}v. \end{aligned}$$

The 2D case: Here, there are two nested summation loops over infinitesimal slices. The instantaneous situation is shown in figure L.1. Applying the rule of integration by parts first to the inner loop yields:

$$\iint\limits_{\Omega} \phi \frac{\partial \psi}{\partial x} \,\mathrm{d}x \,\mathrm{d}y = \int_{y_B}^{y_T} \left(\phi\psi\big|_{x=x_R} - \phi\psi\big|_{x=x_L} \right) \mathrm{d}y - \iint\limits_{\Omega} \frac{\partial \phi}{\partial x} \psi \,\mathrm{d}x \,\mathrm{d}y$$

The first term on the right can be further modified by finding a way of writing $\mathrm{d}y$ differently. This can be done with the normal vector $\hat{n}$ of $\mathrm{d}\Gamma$. The normal vector is $\hat{n} = \binom{n_x}{n_y} = \binom{\cos\alpha}{\sin\alpha}$ if α is the angle to the x-axis. The geometric situation on the right end of the shaded stripe in figure L.1 allows to infer $n_x = \cos\alpha = \frac{\mathrm{d}y}{\mathrm{d}\Gamma}$, whereas on the left $-n_x = -\cos\alpha = \frac{\mathrm{d}y}{\mathrm{d}\Gamma}$. Replacing $\mathrm{d}y$ with $-n_x \mathrm{d}\Gamma$ on the left and with $n_x \mathrm{d}\Gamma$ on the right, the term can be modified according to

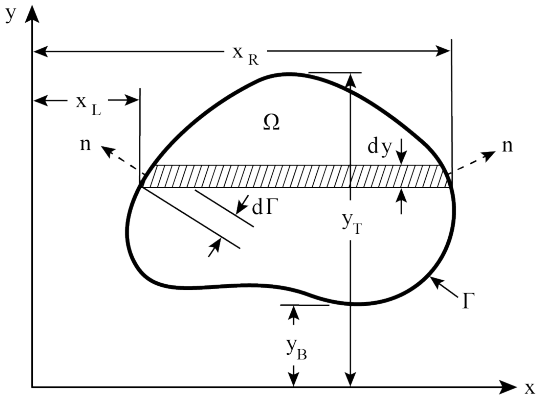


Figure L.1 Definitions for integration in two dimensions.

$$\begin{aligned}\int_{y_B}^{y_T} \left(\phi\psi\big|_{x=x_R} - \phi\psi\big|_{x=x_L}\right) \mathrm{d}y &= \int_{y_B}^{y_T} \phi\psi\big|_{x=x_R}\, n_x \,\mathrm{d}\Gamma + \int_{y_B}^{y_T} \phi\psi\big|_{x=x_L}\, n_x \,\mathrm{d}\Gamma \\ &= \oint_\Gamma \phi\psi\, n_x \,\mathrm{d}\Gamma.\end{aligned}$$

But why is the last line possible? This is because the closed integral around one slice like the shaded one has four sides, but when concatenating the closed integrals around all slices, the long top and bottom legs cancel each other and the closed integral along $\mathrm{d}\Gamma$ can be made just of the side pieces. Therefore we have

$$\iint_\Omega \phi\frac{\partial\psi}{\partial x}\,\mathrm{d}x\,\mathrm{d}y = -\iint_\Omega \frac{\partial\phi}{\partial x}\psi\,\mathrm{d}x\,\mathrm{d}y + \oint_\Gamma \phi\psi\, n_x\,\mathrm{d}\Gamma$$

Similarly, if the partial derivatives are with respect to y, one can write

$$\iint_\Omega \phi\frac{\partial\psi}{\partial y}\,\mathrm{d}x\,\mathrm{d}y = -\iint_\Omega \frac{\partial\phi}{\partial y}\psi\,\mathrm{d}x\,\mathrm{d}y + \oint_\Gamma \phi\psi\, n_y\,\mathrm{d}\Gamma.$$

The 3D case: A similar procedure in this case yields

$$\iiint_\Omega \phi\frac{\partial\psi}{\partial y}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = -\iiint_\Omega \frac{\partial\phi}{\partial y}\psi\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z + \oint_\Gamma \phi\psi\, n_y\,\mathrm{d}\Gamma.$$

From scalar to vector fields: Now imagine the function ψ is a vector field, i. e. a list containing in each component a scalar-valued function for which the above identities can be applied. Taking the above equation and modifying the naming convention by replacing every ϕ with ϕ_i is pure writing cosmetic:

$$\iiint_\Omega \phi_i\frac{\partial\psi}{\partial x_j}\,\mathrm{d}\Omega = -\iiint_\Omega \frac{\partial\phi_i}{\partial x_j}\psi\,\mathrm{d}\Omega + \oint_\Gamma \phi_i\psi\, n_j\,\mathrm{d}\Gamma.$$

The meaning didn't change at all, as both labels, the old ϕ and the new ϕ_i, stand for functions projecting $\mathbb{R}^3$ onto $\mathbb{R}$. But by saying that this holds $\forall i$ and that ϕ_i is one

component of a function $\vec{\phi}(\vec{x}) : \mathbb{R}^3 \to \mathbb{R}^3$, and by summing three such equations where the partial derivatives are chosen so j matches i, it becomes a statement that can be written in the following way:

$$\iiint_\Omega \left(\sum_i \phi_i \frac{\partial \psi}{\partial x_i} \right) \mathrm{d}\Omega = - \iiint_\Omega \left(\sum_i \frac{\partial \phi_i}{\partial x_i} \right) \psi \, \mathrm{d}\Omega + \oint_\Gamma \left(\sum_i \phi_i n_i \right) \psi \, \mathrm{d}\Gamma$$

And this is nothing else than

$$\iiint_\Omega \vec{\phi} \cdot \vec{\nabla} \psi \, \mathrm{d}\Omega = - \iiint_\Omega (\vec{\nabla} \cdot \vec{\phi}) \psi \, \mathrm{d}\Omega + \oint_\Gamma (\vec{\phi} \cdot \hat{n}) \psi \, \mathrm{d}\Gamma.$$

This equation can be used for two important theorems.

Divergence theorem: Setting $\psi = 1$ reduces the above equation to the divergence theorem:

$$\iiint_\Omega \vec{\nabla} \cdot \vec{\phi} \, \mathrm{d}\Omega = \oiint_\Gamma \vec{\phi} \cdot \hat{n} \, \mathrm{d}\Gamma.$$

Product of gradients: Setting $\vec{\phi} = \vec{\nabla} \chi$ modifies the theorem so it can be used for treating the dot product of two gradients:

$$\int_\Omega \vec{\nabla} \chi \cdot \vec{\nabla} \psi \, \mathrm{d}\Omega = - \int_\Omega (\vec{\nabla} \cdot \vec{\nabla} \chi) \psi \, \mathrm{d}\Omega + \oint_\Gamma \psi \vec{\nabla} \chi \cdot \hat{n} \, \mathrm{d}\Gamma.$$

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
$\hat{n}$	unit vector in the normal direction
u, v	generic functions
x, y, z	spatial coordinates

List of Greek quantity symbols

Symbol	Description
α	generic angle
Γ	boundary of the region/domain under consideration
ϕ, χ, ψ	generic functions
Ω	volume, i. e. spatial region/domain under consideration

Appendix M

Some basic definitions of solid mechanics

(This description is largely based on the instructive script by Kuhl & Meschke of the Ruhr University Bochum [244].)

Be $\boldsymbol{u}$ the displacement field $\boldsymbol{u} = \vec{u}(\vec{x}) : \mathbb{R}^3 \to \mathbb{R}^3$ describing the deformation state of a piece of solid matter. The Green-Lagrange strain tensor (finite strain) is defined as

$$\boldsymbol{\mathcal{E}} = \frac{1}{2}\left(\nabla\boldsymbol{u} + \nabla^T\boldsymbol{u} + \nabla^T\boldsymbol{u}\cdot\nabla\boldsymbol{u}\right).$$

The displacement gradient can be decomposed into a symmetric and a skew-symmetric part:

$$\nabla\boldsymbol{u} = \nabla^{\mathrm{sym}}\boldsymbol{u} + \nabla^{\mathrm{skw}}\boldsymbol{u} = \frac{1}{2}(\nabla\boldsymbol{u} + \nabla^T\boldsymbol{u}) + \frac{1}{2}(\nabla\boldsymbol{u} - \nabla^T\boldsymbol{u})$$

which allows the shorter writing of the finite strain

$$\boldsymbol{\mathcal{E}} = \nabla^{\mathrm{sym}}\boldsymbol{u} + \frac{1}{2}\nabla^T\boldsymbol{u}\cdot\nabla\boldsymbol{u}$$

where the two summands mark the separation into a first linear term and a second term nonlinear in $\boldsymbol{u}$. The infinitesimal strain tensor $\boldsymbol{\varepsilon}$ is defined in terms of a geometrically linear theory of small strains as

$$\boldsymbol{\varepsilon} = \nabla^{\mathrm{sym}}\boldsymbol{u}$$

which has the shape

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{pmatrix} = \begin{pmatrix} \partial_1\boldsymbol{u}_1 & \frac{1}{2}(\partial_1\boldsymbol{u}_2 + \partial_2\boldsymbol{u}_1) & \frac{1}{2}(\partial_1\boldsymbol{u}_3 + \partial_3\boldsymbol{u}_1) \\ \frac{1}{2}(\partial_2\boldsymbol{u}_1 + \partial_1\boldsymbol{u}_2) & \partial_2\boldsymbol{u}_2 & \frac{1}{2}(\partial_2\boldsymbol{u}_3 + \partial_3\boldsymbol{u}_2) \\ \frac{1}{2}(\partial_3\boldsymbol{u}_1 + \partial_1\boldsymbol{u}_3) & \frac{1}{2}(\partial_3\boldsymbol{u}_2 + \partial_2\boldsymbol{u}_3) & \partial_3\boldsymbol{u}_3 \end{pmatrix}$$

and the property $\varepsilon_{ij} = \varepsilon_{ji}$ for all off-diagonal elements. Because of the symmetry it suffices to give just six of the nine elements and a convention has been established allowing to write $\boldsymbol{\varepsilon}$ as a vector

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{22} & \varepsilon_{33} & 2\varepsilon_{12} & 2\varepsilon_{23} & 2\varepsilon_{13} \end{pmatrix}^T$$

where the factor 2 is of relevance because it allows to express the internal strain energy as a product of strain and stress in the tensor notation $\boldsymbol{\varepsilon} : \boldsymbol{\sigma}$ as well as in the shortened vector notation $\boldsymbol{\varepsilon} \cdot \boldsymbol{\sigma}$. In order to express the defining equation for the strain in the vector notation, an extra differential operator needs to be defined:

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ 2\varepsilon_{12} \\ 2\varepsilon_{23} \\ 2\varepsilon_{13} \end{pmatrix} = \begin{pmatrix} \partial_1 & 0 & \\ 0 & \partial_2 & 0 \\ 0 & 0 & \partial_3 \\ \partial_2 & \partial_1 & 0 \\ 0 & \partial_3 & \partial_2 \\ \partial_3 & 0 & \partial_1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \mathbf{D}_\varepsilon \boldsymbol{u}$$

.

The Cauchy stress tensor is defined as to yield the ratio of force $\Delta \boldsymbol{F}$ to area ΔA for arbitrarily oriented flat test surfaces in the limit of small areas $\Delta A \to 0$:

$$\boldsymbol{T}^{(n)} = \lim_{\Delta A \to 0} \frac{\Delta \boldsymbol{F}}{\Delta A} = \boldsymbol{\sigma} \cdot \boldsymbol{n}$$

where $\boldsymbol{n}$ is the normal vector on the test surface. $\boldsymbol{\sigma}$ is again a symmetric 3×3-matrix and the convention to write it in a pseudovector format is

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{22} & \sigma_{33} & \sigma_{12} & \sigma_{23} & \sigma_{13} \end{pmatrix}^T = \begin{pmatrix} \sigma_{11} & \sigma_{22} & \sigma_{33} & \tau_{12} & \tau_{23} & \tau_{13} \end{pmatrix}^T$$

where the shear stress components are denoted as τ_{ij}. With these definitions it is possible to write down the local force equilibrium or momentum balance for solid matter:

$$\varrho \ddot{\boldsymbol{u}} = \operatorname{div} \boldsymbol{\sigma} + \varrho \boldsymbol{b} = (\partial_j \sigma_{ij} + \varrho b_i) \boldsymbol{e}_i$$

with

$$\operatorname{div} \boldsymbol{\sigma} = \begin{pmatrix} \partial_1 \sigma_{11} + \partial_2 \sigma_{12} + \partial_3 \sigma_{13} \\ \partial_1 \sigma_{21} + \partial_2 \sigma_{22} + \partial_3 \sigma_{23} \\ \partial_1 \sigma_{31} + \partial_2 \sigma_{32} + \partial_3 \sigma_{33} \end{pmatrix}.$$

In order to be able to write this also in the pseudovector notation it is necessary to define another differential operator $\mathbf{D}_\sigma = \mathbf{D}_\varepsilon^T$, so

$$\varrho \ddot{\boldsymbol{u}} = \mathbf{D}_\sigma \boldsymbol{\sigma} + \varrho \boldsymbol{b} = (\partial_j \sigma_{ij} + \varrho b_i) \boldsymbol{e}_i.$$

When a linear string is elongated from rest then the force builds up according to $F = -kx$ (Hooke's law) and the work done on the spring is $W = \int_{x_i=0}^{x_f > x_i} F(x)\,\mathrm{d}x = [-\frac{1}{2}kx^2]_{x_i=0}^{x_f}$. Thus, the work received by the spring, the stored energy, is $U = -W = \frac{1}{2}kx_f^2$. The spring constant $k = \frac{\mathrm{d}F}{\mathrm{d}x} = \frac{\mathrm{d}^2U}{\mathrm{d}x^2}$ is the quantity characterising the linear spring, it is the knowledge allowing the prediction of forces and energies at arbitrary values of x. Analogously, one wants to define a similar symbol giving the characteristics of an elastic material: this is the constitutive tensor $\boldsymbol{\mathcal{C}}$ defined through

$$\boldsymbol{\mathcal{C}} = \frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{\varepsilon}} = \frac{\partial^2 U}{\partial \boldsymbol{\varepsilon} \otimes \partial \boldsymbol{\varepsilon}} = \mathcal{C}_{ijkl}\, \boldsymbol{e}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_k \otimes \boldsymbol{e}_l$$

which is a highly symmetric fourth order tensor with $\mathcal{C}_{ijkl} = \mathcal{C}_{jikl} = \mathcal{C}_{jilk} = \mathcal{C}_{ijlk}$. If $\boldsymbol{\mathcal{C}}$ is independent of the strain, then one deals with a physically linear (or material linear) constitutive law and one can write for the energy [516]

$$U = \frac{1}{2} \sum_{ijkl} \mathcal{C}_{ijkl} \varepsilon_{ij} \varepsilon_{kl}$$

and for the stress

$$\sigma_{ij} = \frac{\partial U}{\partial \varepsilon_{ij}} = \sum_{k,l} \mathcal{C}_{ijkl} \varepsilon_{kl}$$

or just

$$\boldsymbol{\sigma} = \boldsymbol{\mathcal{C}} \boldsymbol{\varepsilon}$$

which is the generalised Hooke's law. $\boldsymbol{\mathcal{C}}$ with its four indices has $3^4 = 81$ entries. The symmetry among many of its $3^4 = 81$ entries stems from the symmetry of the 3×3 matrices it must connect. The pseudovector notation, where $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$ are vectors with 6 independent components, can be kept up, and $\boldsymbol{\mathcal{C}}$ must then be written as a 6×6 matrix denoted here as $\boldsymbol{C}$. That this matrix must again be symmetric means that the most general material law can be defined by a maximum of 21 constants. Wilson [516] notes that symmetry with respect to a plane reduced the number to 13, symmetry about three mutually perpendicular planes reduced it to 9, in cubic crystals it narrowed down to three independent constants, in an isotropic solid to two, and in an ideal fluid to one. For a physically linear isotropic solid, the two independent numbers can be put in relation with the macroscopically measurable properties Young's modulus E (modulus of elasticity), Poisson's ratio ν, and shear modulus G. Since only two of these three are independent, there are connecting equations which can be put in a simple form by defining Lamé constants μ and λ according to

$$\begin{aligned} \mu &= \frac{E}{2(1+\nu)} = G \quad \text{and} \\ \lambda &= \frac{\nu E}{(1+\nu)(1-2\nu)}. \end{aligned}$$

With these one can write the constitutive law for the linear isotropic case as

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{23} \\ \sigma_{13} \end{pmatrix} = \begin{pmatrix} 2\mu+\lambda & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & 2\mu+\lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & 2\mu+\lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix} \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ 2\varepsilon_{12} \\ 2\varepsilon_{23} \\ 2\varepsilon_{13} \end{pmatrix} = \boldsymbol{C} \boldsymbol{\varepsilon},$$

and the inverse relation can be written as

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \varepsilon_{33} \\ 2\varepsilon_{12} \\ 2\varepsilon_{23} \\ 2\varepsilon_{13} \end{pmatrix} = \begin{pmatrix} 1 & -\nu & -\nu & 0 & 0 & 0 \\ -\nu & 1 & -\nu & 0 & 0 & 0 \\ -\nu & -\nu & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{E}{G} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{E}{G} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{E}{G} \end{pmatrix} \begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{23} \\ \sigma_{13} \end{pmatrix} = \boldsymbol{C}^{-1} \boldsymbol{\sigma}.$$

The matrix $\boldsymbol{C}^{-1}$ is called the compliance matrix or the material's elastic compliance. In English literature it is often denoted as $\boldsymbol{S}$. For specifying piezoelectric ceramics, the entries of $\boldsymbol{C}$ and $\boldsymbol{S}$ are given either at constant E-field (then denoted with c_{ij}^E and s_{ij}^E) or at constant D-field (c_{ij}^D, s_{ij}^D).

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
A	surface area
$\boldsymbol{b}$	load force field
$\boldsymbol{C}$	elasticity tensor (constitutive tensor) in pseudovector notation
$\boldsymbol{\mathcal{C}}$	elasticity tensor (constitutive tensor)
c_{ij}^E, c_{ij}^D	elasticity tensor entries given at constant field strength
$\mathbf{D}_\varepsilon, \mathbf{D}_\sigma$	differential operators
E	Young's modulus
$\boldsymbol{\mathcal{E}}$	Green-Lagrange strain tensor (finite strain)
$\boldsymbol{e}_i$	unit vector
$F, \boldsymbol{F}$	force
G	shear modulus
k	spring constant, stiffness
$\boldsymbol{n}$	unit vector in the normal direction
$\mathbb{R}$	real numbers
$\boldsymbol{S}$	compliance tensor
s_{ij}^E, s_{ij}^D	compliance tensor entries given at constant field strength
$\boldsymbol{T}^{(\boldsymbol{n})}$	Cauchy stress tensor
U	internal enery
$u, \boldsymbol{u}, \vec{u}$	displacement
W	work
$x, \vec{x}$	distance, spatial coordinate

List of Greek quantity symbols

Symbol	Description
$\boldsymbol{\varepsilon}$	strain tensor
λ	first Lamé parameter
μ	second Lamé parameter, equates to shear modulus
ν	Poisson's ratio
ϱ	density
$\boldsymbol{\sigma}$	stress tensor
τ_{ij}	shear stress components

Appendix N

Some more basics on damping

N.1 The concept of receptance

If a harmonic force $f(t) = Fe^{i\omega t}$ (with $F \in \mathbb{R}$, the amplitude) acts on a system and the system reacts by exhibiting the displacement $x(t) = Xe^{i\omega t}$ varying with the same frequency ω at the point of application of the force, then a proportionality constant α can be defined for the system, if its equations of motion are linear, connecting the response with the excitation by

$$x = \alpha F e^{i\omega t}.$$

The constant $\alpha = \alpha(\omega)$ is called 'the direct receptance at x'. It is frequency-dependent, but independent of the excitation amplitude F. Let us consider for example a mass being able to move frictionless along one dimension and subject to the force $f(t) = Fe^{i\omega t}$. Then the displacement response must be a solution to

$$m\ddot{x}(t) = f(t).$$

Assuming one solution might have the form $x(t) = Xe^{i\omega t}$ the equation turns into

$$-m\omega^2 X = F$$

and it can thus be said that

$$\alpha = -\frac{1}{m\omega^2}$$

is the direct receptance at x of the system, because as a consequence of the ansatz X/F equals x/f in this case. Considering as a second elementary example a massless spring with $kx = Fe^{i\omega t}$ one can find that the direct receptance at x is then $\alpha = 1/k$.

Now, let us consider a mass-spring system where

$$m\ddot{x} + kx = Fe^{i\omega t} \tag{N.1}$$

is the equation of motion. Here, the receptance $\alpha = X/F$ turns out to be

$$\alpha = \frac{1}{k - m\omega^2}. \tag{N.2}$$

$\alpha(\omega)$ has a poles at $\omega = \pm\sqrt{k/m}$, this means the resonance frequency ω_0 (sometimes called natural frequency ω_n) is $\sqrt{k/m}$. (For frequencies the reversed sign means nothing else than a phase shift of π.)

N.2 Differential equations and interpretations of damping models

Work is force times distance. In a one-dimensional oscillator as sketched in the insets of figure N.1 energy flows back and forth between the spring and the mass as work is done by one half of the system to the other. For three types of oscillators the figure shows plots of the force acting on the spring depending on the position of the mass. Periodic trajectories in the force-position-plane are lossless if the ways back and forth coincide, whereas an imperfect spring which is not able to recover all the stored energy will split the two paths apart so there is a finite area encircled by the trajectory.[1] The size of the area within is ΔW, the energy dissipated during one oscillation cycle. The figure shows different abstract approaches of modelling spring imperfections which will be motivated and discussed in the following. The models are *dry (Coulomb) friction*, *viscous damping*, and so-called *ideal hysteretic damping*. The corresponding differential equations for a one-dimensional spring-mass system can be written

$$\text{Coulomb:} \quad m\ddot{x} - F_{\text{dry}}\frac{\dot{x}}{|\dot{x}|} + kx = f_{\text{ext}}(t), \quad k \in \mathbb{R}, \tag{N.3}$$

$$\text{viscous:} \quad m\ddot{x} + \gamma\dot{x} + kx = f_{\text{ext}}(t), \quad k \in \mathbb{R}, \tag{N.4}$$

$$\text{id. hyst. A:} \quad m\ddot{x} + \tfrac{h}{\omega}\dot{x} + kx = f_{\text{ext}}(t), \quad k \in \mathbb{R}, \tag{N.5}$$

$$\text{id. hyst. B:} \quad m\ddot{x} + kx = f_{\text{ext}}(t), \quad k = k' + ik'' \in \mathbb{C}, \tag{N.6}$$

$$\text{id. hyst. C:} \quad m\ddot{x} + b\left|\tfrac{x}{\dot{x}}\right|\dot{x} + kx = f_{\text{ext}}(t), \quad k \in \mathbb{R}, \tag{N.7}$$

and with the harmonic ansatz $f_{\text{ext}}(t) = F_{\text{ext}}e^{i\omega t}$ and $x(t) = Xe^{i(\omega t+\varphi)}$ these equations turn into[2]

$$\text{Coulomb:} \quad -m\omega^2 X + i\tfrac{4}{\pi}F_{\text{dry}} + kX = F_{\text{ext}}e^{-i\varphi}, \tag{N.8}$$

$$\text{viscous:} \quad -m\omega^2 X + i\omega\gamma X + kX = F_{\text{ext}}e^{-i\varphi}, \tag{N.9}$$

$$\text{id. hyst. A:} \quad -m\omega^2 X + ihX + kX = F_{\text{ext}}e^{-i\varphi}, \tag{N.10}$$

$$\text{id. hyst. B:} \quad -m\omega^2 X + ik''X + k'X = F_{\text{ext}}e^{-i\varphi}. \tag{N.11}$$

$$\text{id. hyst. C:} \quad -m\omega^2 X + i\tfrac{2b}{\pi}X + kX = F_{\text{ext}}e^{-i\varphi}. \tag{N.12}$$

Viscous damping: That the model of viscous damping is so common has two reasons. On the theoretical side, equation N.4 is a linear differential equation and

[1] The case of areas encircled in clockwise and anti-clockwise manner cancelling each other out is not considered here. Discussing models for loss mechanisms of imperfect springs means all areas are surrounded in clockwise orientation.

[2] The nonlinear differential equation for Coulomb friction is not easily solvable. One option is to replace the nonlinear term $F_{\text{dry}}\,\text{sign}(\dot{x})$ (with the help of the known $x(t)$, e. g. $x(t) = X\sin\omega t$ and $\dot{x}(t) = \omega X\cos\omega t$) by a harmonic series (see [321], p. 60). Here, the viscous damping equation has been used by employing an equivalent damping constant γ. In the viscous case the area in figure N.1 is $\Delta W = \pi\gamma|\omega|X^2$, whereas in the Coulomb friction case the size of the parallelogram-shaped area is $\Delta W = 4XF_{\text{dry}}$. Hence, a frequency-dependent value of $\gamma = \frac{4F_{\text{dry}}}{\pi\omega X}$ has to be used in order to emulate Coulomb friction by equation N.4 which results in equation N.8. This approximation is, however, only valid in the domain of weak damping where realistic solutions $x(t)$ to the two models are still very similar. The same procedure has been used for coming to equation N.12 where according to the friction force term in equation N.7 ΔW equals to $2bX^2$.

it has the simple harmonic solutions shown in appendix K.5. On the experimental side, it is quite easy to implement a velocity-proportional damping mechanism for didactic experiments at laboratory scale. Friction by laminar flow phenomena can be used for this purpose, e. g. by submerging a sidearm of a pendulum in oil, hence the dashpot symbol commonly used for viscous damping elements. The force-over-position trajectory in the middle of figure N.1 is the result of a harmonic movement $x(t) = X \sin \omega t$ and viscous damping. Doubling the frequency ω means doubling the speed everywhere on the trajectory and hence widening the elliptical surface corresponding to the energy loss ΔW by a factor of two. Doubling the amplitude X while keeping the frequency constant means doubling both length and width (because of the necessary velocity increase to cover the larger distance in the same time) of the elliptical surface. That ΔW is proportional to ω and X^2 can also be verified by

$$\Delta W = \oint \gamma \dot{x} \, \mathrm{d}x = \gamma \oint \dot{x}^2 \, \mathrm{d}t = \gamma \int_0^{\frac{2\pi}{\omega}} \omega^2 X^2 \cos^2 \omega t \, \mathrm{d}t = \pi \gamma |\omega| X^2. \tag{N.13}$$

However, the dashpot model of viscous damping is often not suitable to describe damped vibrating systems because when damping is caused by intrinsic material losses, by the imperfect elasticity of solid structural materials, then experimental data on many of the most common construction materials indicates that ΔW is independent of frequency [233].

Coulomb damping. In the Coulombian model of dry friction the friction force between two solid surfaces sliding against each other depends only on the surface properties and the normal force, it is in particular independent of velocity. Adopting such a friction law for the 1D oscillator means that the friction force acting on the mass moving with velocity $\dot{x}$ is $f_{\mathrm{dry}} = -F_{\mathrm{dry}} \operatorname{sign}(\dot{x}) = -F_{\mathrm{dry}} \frac{\dot{x}}{|\dot{x}|}$. This term renders the equation of motion N.3 nonlinear and it leads to the force-position trajectories plotted in the top diagram of figure N.1. That solutions $x(t)$ to equation N.3 are not as well-behaved as the ones of N.4 can be realised by looking at what happens as a response to harmonic excitations at very low frequencies. In that regime inertia plays no role, the mass follows the excitation force in phase, and because there is a jump in the friction force when $\dot{x}$ switches its sign the mass will stop for finite times at the outside positions before returning. At higher frequencies the influence of the friction force jumps shrinks against the inertia effects, but it never disappears, hence the assumption of a sinusoidal $x(t)$ always remains just an approximation. The energy dissipated per cycle is very easy to calculate for the Coulomb friction model because ΔW is just the size of the parallelogram-shaped area in the top diagram of figure N.1 which is $\Delta W = 4 X F_{\mathrm{dry}}$. Besides the nonlinearity, the fact that ΔW is not proportional to X^2 is the second drawback of the Coulomb friction model, because it is thus not suitable to describe the experimental data on structural damping.

Ideal hysteretic damping: The model of ideal hysteretic damping is often equated with the concept of a complex stiffness $k = k' + ik''$ (as in eq. N.10) or a complex Young's modulus $E = E' + iE''$, in which case E' and E'' are called the *storage* and *loss moduli*, respectively [446]. Bishop and Johnson [46] motivate this concept in the following way: replacing γ by h/ω and thereby going from equation N.4 to N.5 is a way to achieve a ΔW independent of frequency and at the same time

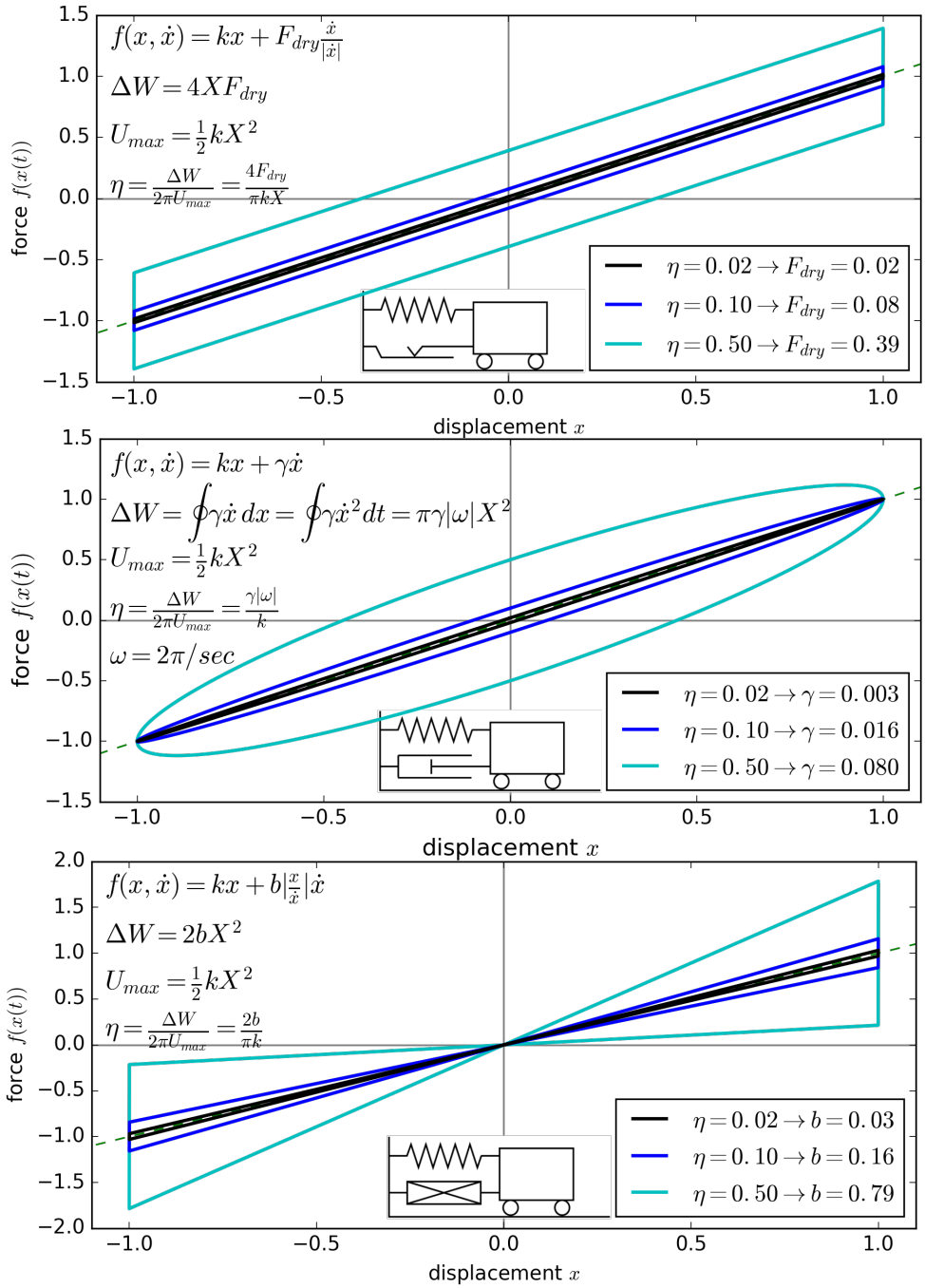


Figure N.1 Hysteresis loops of 1D oscillators with different damping models.
The force exerted by the oscillator mass on the spring-damper element is the sum of the spring force $f_{\text{spring}} = kx$ and the friction force f_{friction}. The plots show stationary oscillator trajectories in the f-x-plane. Work done on (gained from) the spring-damper element corresponds to areas circumscribed in clockwise (anticlockwise) direction. Without damping the oscillator swings back and forth on the straight line $f(x) = kx$ (the green dashed line). But with damping there is always a hysteresis loop in the f-x-plane, and the energy ΔW dissipated per cycle is the area within the loop. In the case of viscous damping (middle) the trajectory in the f-x-plane can in principle take any shape by making special choices on the velocity profile. The present elliptical trajectory is a consequence of the particular case $x(t) = X\sin(\omega t + \varphi)$. The width of the ellipse is proportional to the oscillation velocity and thus ω. Therefore, $\Delta W \propto \omega$, too. In the dry friction (top) and the ideal hysteretic damping model (bottom) the trajectory shape and the encircled area ΔW do not depend on the velocity distribution but only on the displacement amplitude X. These models represent frequency-independent damping. It is clear that the transient trajectories of free oscillators without energy input will be spiralling towards the origin.

proportional to the square relative displacement (or strain) amplitude and thereby correspondence with experimental data. Only the transition into the frequency domain motivates the idea of a complex stiffness $k = k' + ik''$ by collecting all terms on the left hand side of equation N.10 which are of zeroth order in ω and first order in X. The crucial point is that equations N.10 and N.11 are equivalent. That means, in the treatment of forced-harmonic system responses (eq. N.10/N.11) it is mathematically irrelevant whether the starting point has been equation N.5 or N.6. However, not forgetting about the boundary conditions of the motivation for the usage of a complex k is important because equations N.5 and N.6 are not at all equivalent. The latter one implies that a complex-valued force is the result of a real-valued displacement, and this is unphysical [216]. Taking equation N.6 to analyse how a system with ideal hysteretic damping behaves in the time domain means leaving the motivation context according to [46]. Applying the method of Fourier series to find solutions of equation N.6 which are a response to an impulse excitation reveals another feature of that differential equation, namely that it describes a system violating causality, since the solutions start to oscillate already before the excitation happens (exhibiting a so-called *impulse response precursor, IRP*) [92]. The controversy of that interpretation seems to be an interesting issue until now [49]. Nashif et al. [321] note that the quantities k and η in $k = k' + ik'' = k(1 + i\eta)$ should be understood as functions of frequency, $k(\omega)$ and $\eta(\omega)$. They contend that with the help of the settings $k(\omega) = k(-\omega)$ and $\eta(\omega) = -\eta(-\omega)$ in a Fourier series approach one can for the case of a force impulse $f(t) = F\delta(t)$ arrive at an expression for the time response which obeys causality if evalued numerically (see [321], section 2.4.5, p. 78). However, at a more fundamental level it is sometimes pointed out (e. g. [39]) that the term "frequency" is not necessarily clear in the time domain. A problematic situation in that sense is encountered when intending to work with nonconstant functions $k(\omega)$ and $\eta(\omega)$ measured in the frequency domain for constructing solutions to a transient excitation $f(t)$. Which values of k and η should be taken at any given time step?

By contrast, the nonlinear equation N.7 is a physically meaningful differential equation for the time domain which suffices the requirement of ideal hysteretic damping that the friction force is proportional to the strain. For a periodically oscillating displacement history it leads to the closed trajectories depicted at the bottom of figure N.1. The second requirement that ΔW is independent of frequency is met because the friction force is independent of velocity and the shape of the force-position trajectories is independent of the velocity pattern.

The three simple damping models explained above and illustrated in figure N.1 have been defined for the one-dimensional oscillator. But they can also be taken as the basis of a viscoelastic material law.

Viscoelasticity: A material model accounting for energy dissipation during deformation is called *viscoelastic*. Contrary to the meaning of its first part, the term *viscoelasticity* does not only mean the introduction of a laminar viscosity, but rather means in a broader sense any effect with the consequence that the microscopic stress state does not only depend on the strain but also on the deformation history.[3] There

[3] see [479], page 37f and [446], page 10f

exist many more relevant rheological models describing viscoelastic behaviour as the three simplest ones discussed above [479].

N.3 The one-dimensional harmonic oscillator damped by a complex stiffness

(This appendix section follows the nomenclature of [446] and lays out the mathematical steps hinted therein.)

The equation of motion of a forced harmonic oscillator is

$$m\ddot{x} + kx = F_a e^{i\omega t}. \tag{N.14}$$

The obvious difference with respect to equation K.22 is that the velocity-proportional damping term is missing. Here it shall be shown how offsetting k from the real axis can be an alternative way of damping the system, i. e. $k = k' + ik'' = k'(1 + i\eta)$. In order to get the frequency response of the oscillation amplitude, the ansatz

$$x(t) = Ae^{i\omega t} \tag{N.15}$$

will be made where the phase shift is being accounted for by the complex amplitude. This results in

$$\begin{aligned} -mA\omega^2 + kA &= F_a \\ A &= \frac{F_a}{k - m\omega^2} \\ A &= \frac{F_a}{k'(1 + i\eta) - m\omega^2} \end{aligned} \tag{N.16}$$

which can be modified by using the frequency of the undamped resonance $\omega_0 = \sqrt{k'/m}$ into

$$A(\omega) = \frac{F_a/k'}{1 + i\eta - (\frac{\omega}{\omega_0})^2}. \tag{N.17}$$

The magnitude of that frequency response is plotted in figure K.4. It can be seen that unlike viscous damping, stiffness damping does not ensue a shift in the resonance frequency. Evaluating that formula at $\omega = \omega_0$ leads to an expression for the dependency of the resonance peak height on the damping parameter η.

$$|A(\omega{=}\omega_0)| = \frac{F_a}{k'\eta} \tag{N.18}$$

Employing a different ansatz

$$x(t) = |A|e^{i(\omega t + \varphi)} \tag{N.19}$$

allows to get to an equation giving the frequency response of the phase shift φ, since

$$\begin{aligned} -m|A|\omega^2 e^{i(\omega t+\varphi)} + k|A|e^{i(\omega t+\varphi)} &= F_a e^{i\omega t} \\ -m|A|\omega^2 e^{i\varphi} + k|A|e^{i\varphi} &= F_a \\ e^{-i\varphi} &= \frac{|A|}{F_a}(k - m\omega^2) \\ e^{-i\varphi} &= \frac{|A|}{F_a}(k'(1+i\eta) - k'(\tfrac{\omega}{\omega_0})^2) \\ e^{-i\varphi} &= \frac{|A|k'}{F_a}(1 - (\tfrac{\omega}{\omega_0})^2 + i\eta) \;=\; \cos\varphi - i\sin\varphi. \end{aligned} \tag{N.20}$$

Now, $\tan\varphi$ can be determined by dividing the imaginary by the real part

$$\tan\varphi = \frac{\sin\varphi}{\cos\varphi} = \frac{\eta}{(\frac{\omega}{\omega_0})^2 - 1} \tag{N.21}$$

and the phase shift frequency response (plotted in figure K.4) can be written as

$$\varphi(\eta,\omega) = \arctan\left(\frac{\eta}{(\frac{\omega}{\omega_0})^2 - 1}\right). \tag{N.22}$$

Comparing equation N.22 to the angular response of the viscously damped oscillator (eq. K.28)

$$\begin{aligned} \varphi_{\text{visc}}(\gamma,\omega) &= \arctan\left(\frac{\omega\gamma}{m\omega^2 - k}\right) \\ &= \arctan\left(\frac{\omega\gamma/k}{(\frac{\omega}{\omega_0})^2 - 1}\right) \end{aligned} \tag{N.23}$$

reveals again that the behaviour caused by the complex stiffness can be expressed in terms of viscous damping with a frequency-dependent equivalent damping constant $\gamma = \eta k/\omega$.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
A	ansatz parameter
b	alternative damping coefficient/factor (hysteretic damping)
E	Young's modulus
F	force amplitude
f	force
h	imaginary part of stiffness (hysteretic damping)
i	unit value of imaginary numbers
k	spring constant, stiffness
m	mass

$\mathbb{R}$	real numbers
t	time
U	internal enery
W	work
X	amplitude (of oscillator motion)
x	distance, spatial coordinate

List of Greek quantity symbols

Symbol	**Description**
α	direct receptance
γ	damping coefficient/factor (viscous damping)
$\delta(t)$	infinitesimally narrow impulse function
η	loss factor
ω	angular frequency
ϕ	phase angle

Appendix O

Experimental Characterisation of sonofusion resonators

The core equipment of any sonoluminescence or sonofusion experiment is the acoustic resonator. On the one hand this is just a vessel holding the liquid in place, so that the geometric shape of its volume allows the liquid to oscillate in a useful mode shape and frequency. On the other hand the vessel has to be able to vibrate itself, so that it can carry the vibration generator on the outside and transmit its movements to the liquid on the inside. In our case the vessel is made of glass and the vibration generator is a hollow cylinder of piezoelectric ceramic glued around the vessel. The piezo ring is acting as the electromechanic transducer transforming electric energy into mechanic deformation or movement. If geometrical shapes, masses, and stiffnesses of all three parts, the liquid, the vessel, and the transducer are tuned to fit, and the damping is low, then it is possible to let the liquid oscillate strong enough to show acoustic cavitation. For sonofusion experiments in particular, we assume that the sound pressure in the liquid should become as high as possible, and the existing resonator designs were built with that goal. An experimental characterisation of the resonators by recording acoustic and electric properties of the resonators is necessary to evaluate whether they behave as intended. This appendix chapter describes the corresponding lab setup which was perfected during the measurement campaign at RPI.

O.1 Instrumentation for controlling and characterising a resonator

For characterisation campaigns the acoustic resonator had been embedded in a web of multiple devices forming several signal processing chains. There was one chain leading towards the resonator for the driving voltage signal and several chains leading away from the resonator for carrying and processing the diverse sorts of measurement data to be recorded. A LabVIEW® program[1] had been developed which controlled variable settings of many of the involved devices and could automate a substantial

[1] A list of all used Labview applications with short descriptions can be found in appendix P.1.

Table O.1 List of electronic devices involved in resonator driving and characterisation. NI is short for National Instruments® and HP for Hewlett-Packard®.

short name	brand & type	description	GPIB[2]
lab PC	standard PC	OS: Windows®-NT	control
ADC card	NI PCI-6025E	analogue-to-digital converter inside the PC	
BNC adaptor unit	NI BNC-2090	offers access for BNC cables into the ADC card	
function generator	HP 33120A	15 MHz synthesised function generator	yes
amplifier	Wheelock® AA-250	single-channel amplifier with outputs 25 V (2.5 Ω) and 70 V (20 Ω)	
multimeter	HP 34401A	digital multimeter	yes
scope A	HP 54603B	2-channel oscilloscope with 60 MHz bandwidth	yes
scope B	HP 54615B	2-channel oscilloscope with 500 MHz bandwidth	yes
scope C	Tektronix® 2235	simple oscilloscope with no data processing utilities	
transformer	in-house by RPI team [250]	based on toroidal ferrite core, undocumented	
hydrophone	PCB Piezotronics® model S113-A26	steel encapsulated piezo pressure probe	
signal conditioner	PCB Piezotronics® model 482A21	signal conditioner for hydrophone	
current probe	Tektronix® AM 503	current probe (coil surrounding cable) and amplifier	
displacement probe	Ortofon® OMB 5E	moving magnet turntable pickup needle	
phono amp	Realistic® STA-530	old-school stereo HiFi amplifier	

part of the measurement campaigns. Table O.1 lists the devices often referred to in the following section.

O.1.1 The lab PC and its software and periphery

The lab PC served as central computer workstation integrating the measurement device network. It hosted the program LabVIEW® and was used to control peripheric devices and collect data for storage. GPIB[2] data bus cables were used for controlling the function generator and the oscilloscopes. For analogue-to-digital conversion (ADC) of the acoustic signals, a PCI-6025E data-acquisition board from National Instruments® (NI), mounted on a PCI slot inside the lab PC, was used. It received signals through a BNC-2090 adaptor unit (also by NI) placed outside the computer. A collection of proprietary and in-house Labview codes[1] was used and extended for various resonator driving an measurement situations.

[2]GPIB is short for "General Purpose Interface Bus" or "General Purpose Instrumentation Bus", they are all terms describing connection hardware and data transfer protocols following the IEEE-488 standard.

O.1.2 Generating the voltage signal feeding the piezoelectric transducer

The first element of the voltage supply chain was a function generator of type HP-33120A by Hewlett-Packard® producing a low-voltage sine wave. The amplitude had been generally kept at[3] 100 mV_{pp} and the frequency was a state variable controllable via GPIB. The signal was amplified in the next element of the chain, using a Wheelock® AA-250 amplifier and its 70 V/20 Ω output channel. Lastly, a transformer[4] stepped up the voltage and this signal was fed directly to the piezoelectric transducer glued to the glass resonator.

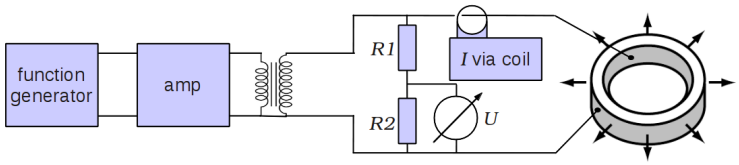


Figure O.1 The circuitry supplying the driving voltage to the piezo ring.
The resistors R_1 and R_2 were labelled 100 kΩ and 1 kΩ with a manufacturer precision rating of ±2% and were measured by multimeter with (99.45 ± 0.05) kΩ and (999.1 ± 0.1) Ω, respectively.

O.1.3 Setup for capturing electrical characterisation signals

The circuitry for picking up the sinusoidal signals of transducer driving voltage and current was connected to the voltage supply chain as shown in figure O.1. Since the driving voltage for the piezo transducer was in the order of 1000 V, which is too much for the ADC card, a reduced voltage signal was captured by feeding the voltage drop across the resistor R_2 into the ADC. A corresponding correction factor in the Labview code reconstructed the correct amplitude values for the voltage signal before being stored (see appendix P.1.3 for details). After digitising the oscillating signal, knowledge about its amplitude and phase is desired. Figure P.2 in appendix P.1.3 shows how this was accomplished with usage of common Labview subroutines.

The current flowing in and out of the piezo ring's electrode surfaces was measured through an inductive method instead of a voltage drop across a low-ohmic resistor. A Tektronix® AM 503 current probing coil and amplifier set was used for that purpose (see appendix P.1.3 for details).

O.1.4 Setup for capturing sound pressure

A PCB Piezotronics® model S113-A26 hydrophone was used to probe sound pressure signals within acetone-filled resonators. This type of hydrophone consists of a small piezoelectric transducer and a small amplifier housed in a pencil-shaped steel tube

[3] The subscript "pp" means peak-to-peak.

[4] Transformer: toroidal ferrite core (OD ≈ 7 cm) with copper wire windings, manufactured in-house at RPI as part of the project of Saglime et al. [250, 390]; it had been designed to work at 20 kHz and therefore impedance-matched with the PZT ring; according to figure 26 on page 35 in [250] (or figure 28 on page 49 in [390]) its windings have a ratio of 1:10. Having examined it with a sensor coil of 8 windings at 20 kHz (and 150 kHz) resulted in rough estimations of 24 (16) windings for the primary coil and 192 (160) for the secondary one.

of diameter 6.3 mm. The flat front surface forms the sensitive pressure probe. A PCB Piezotronics® model 482A21 signal conditioner was necessary to supply the constant current excitation driving the transducer amplifier within the hydrophone housing and for decoupling the output signal from the DC bias signal [305]. The calibration table for converting the voltage output from the signal conditioner into sound pressure is given in appendix P.2.

The hydrophone was kept in place along the central axis of the resonators by a special aluminium flange. This flange looked exactly like the top cover part of resonator Nº 8 (visible in figure I.5) with the only difference of having an extra outlet for the hydrophone in its centre. In that outlet the hydrophone could be fixed and sealed by a rubber O-ring squeezed by a nut. That simple rubber ring fixation also allowed to push and pull the hydrophone into different vertical positions. This hydrophone positioning flange could be used with both resonators, chambers 5 and 8. Its main purpose was to allow constant degassing at the liquid's vapour pressure even with the top piston removed and the hydrophone in place. In some cases, e.g. for very low hydrophone positions when the length of the pencil-shaped hydrophone was not long enough to reach from the sealing ring down to the desired location, the hydrophone was just hung by its cable. During such measurements the acetone surface was in directly exposed to the lab atmosphere. The limiting factor for sound pressure amplitude measurements is cavitation on the probe itself. The cavitation strength of acetone on a steel surface is considerably lower than in the bulk. And it is further lowered once air gets dissolved in the liquid over time.

O.1.5 The wall microphones

Microphones made of small pills of piezoceramic glued on the outside of the resonator are part of most cavitation experiments. They are perfectly suited to pick up the high-frequency noise caused by the implosion of cavitation bubbles. However, interpreting the acoustic properties of the resonator based on their signals is problematic due to the reasons explained below.

Pill microphones were mounted on the outside of the resonator glass walls. They are small round discs of piezoelectric ceramic polarised parallel to the rotation axis. The same type C54000 (with diameter 0.252″ (6.4 mm) and thickness 0.080″ (2.03 mm)) from Channel Industries® was used as in the project by Saglime et al. [250, 390]. After soldering thin connection wires to the silver electrodes on both sides they were glued to the resonator glass walls with a drop of two-component epoxy big enough to entirely coat one side of the disc and connect it to the glass surface, which also implies that the whole solder joint must be immersed. It is the acceleration force parallel to its polarisation of the pill microphone sitting on the oscillating glass surface that is responsible for a deformation of the little disc which translates into a time-varying voltage signal that can be measured across the electrodes. Unfortunately, not only this signal was being picked up by the microphone's connection cables. When the pill microphone was in close proximity to the large piezo ring driving the resonator, which was the case in all resonators examined here, then the electric stray field around the piezo ring induced an additive signal in the microphone cables. These cables therefore had to be wound to a tight helix. Ideally,

coaxial cables should have been used in order to minimise the induced signal, but this would also not have prevented the necessity of the two cable leads opening up near the microphone and going different paths to reach the two separate electrodes. A measurement of this undesired "antenna effect" of the pill microphones is shown in figure O.2. Another point is that once the microphone is glued to the glass surface, no control measurements without contact to the glass surface can be made any more. The susceptibility for stray fields, the invariable position on the glass wall, and the relatively large mass (the microphones are of similar thickness as the glass wall) influencing the glass wall movement are all reasons why the pill microphones are no ideal means to quantitatively examine the normal displacement of the glass wall.[5] But they are perfectly fine to trace cavitation events because that ability relies on the detection of sparks of noise generated by violent bubble collapses covering a much higher frequency range than the resonator's excitation frequency. Saglime et al. [250, 390] used a high-pass filter with a cutoff frequency of about 230 kHz to isolate the cavitation traces from all the lower-frequency signals.

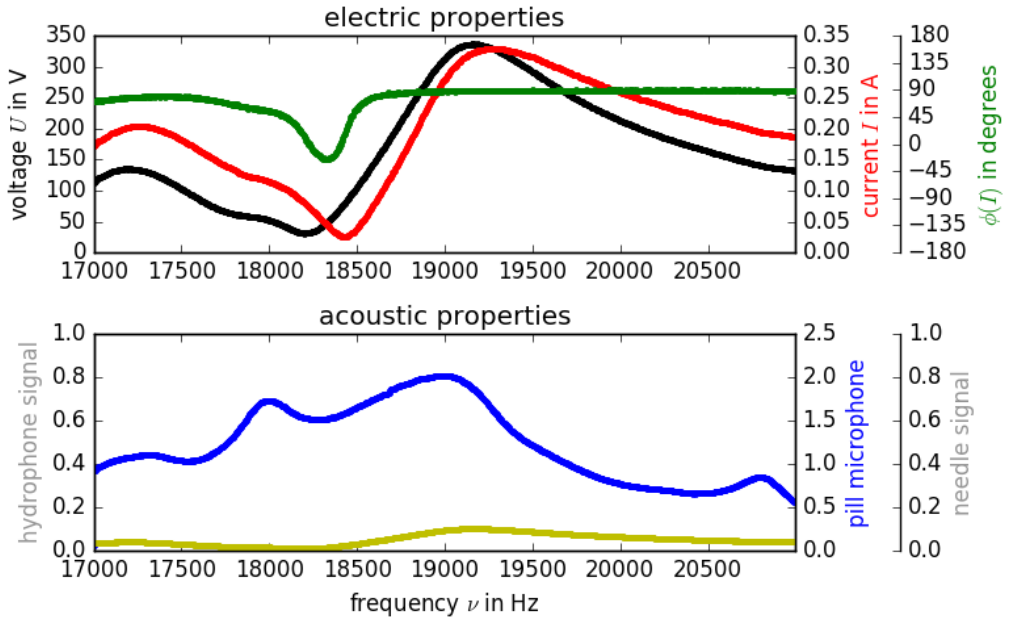


Figure O.2 Susceptibility of the microphone signal towards the electric stray field of the transducer.
In this recording (case 35) an additional pill microphone was placed in close proximity (few millimetres) to the transducer's bottom rim where the stray field is strongest, but it was not in physical contact with the resonator. While the blue curve shows the pill microphone glued on the glass wall, the signal from the stray field probe is plotted in yellow. It can be seen that the latter one picks up a voltage-proportional signal. Its amplitude is up to 15 % of the one of the glued-on microphone at the resonance peaks, and the ratio goes up to 20 % in between the peaks. The effect on the glued-on microphones is assumed to be weaker, as they are all more than a centimetre away from the transducer.

[5] Still another reason is that the spring constant of a pill microphone's epoxy fixation on the glass depends on the geometry of the epoxy drop and the embedded solder joint resulting in a different transfer function of each microphone. So, each microphone would in principle require an own calibration.

O.1.6 Setup for capturing wall displacement

A simple system for probing the displacement of the resonator's outside surfaces was implemented based on the use of a turntable pickup cartridge. The particular advantage this type of measurement has over the hydrophone and the glued-on pill microphones for generating validation data for FEM simulations is that its probing position can be easily changed and such position changes do not affect the acoustic properties of the resonator.

The pickup needle: A standard record player-type pickup needle, model Ortofon® OMB 5E was used. The cartridge contains a diamond tip on a stylus in connection with a transducer of type moving-magnet (MM). The back side of the pickup needle cartridge features four pins for the two stereo channels "left" and "right". During experimentation with and implementation of the circuitry setup, the problem became apparent that the strong electric stray field emanating from the piezo ring induced undesired additional signals into the wiring of the displacement measurement setup, just like in the case of the pill microphones. But here the problem could be solved by using high-quality coaxial cables, by implementing a careful shielding of the whole cartridge-to-coax cable connection area where the coaxial conductor geometry must be broken up, and because the transducer part inside the cartridge is factory-equipped with proper shielding. Figure O.3 shows that there is no residual signal left after detaching the needle from the glass surface.

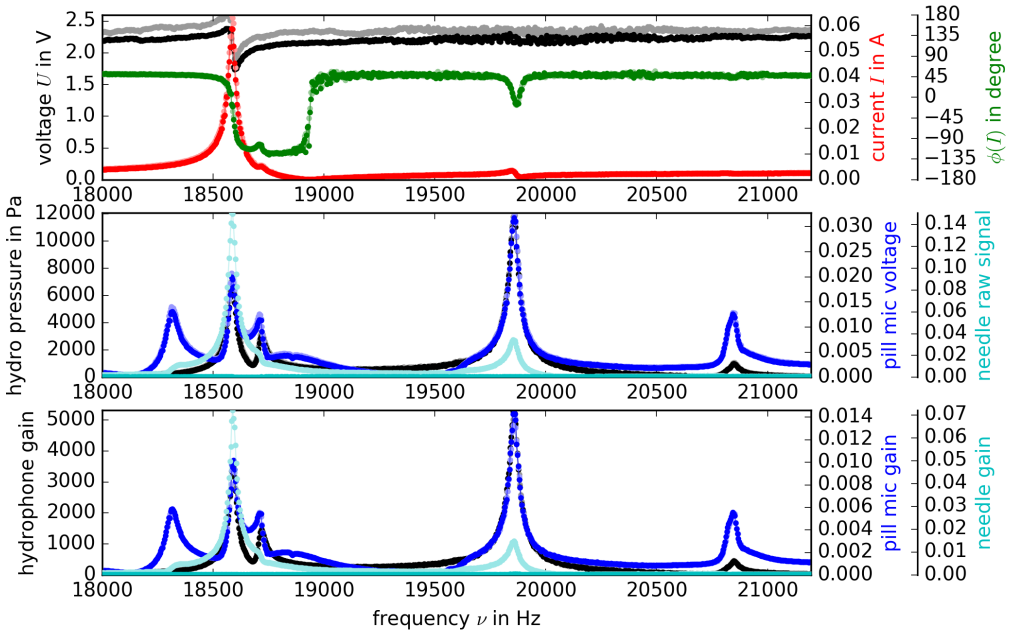


Figure O.3 Stray field susceptibility check on the pickup needle. This plot shows that the signal from the pickup needle does not suffer from being influenced by the electric stray field from the transducer electrodes as a result of proper shielding. The data sets 252 (shaded) & 253 (full colours) are compared. In both cases the vertical position of the needle was 3 cm above the piezo ring. The only difference of the second recording with respect to the first is that the needle cartridge had been pulled back from the glass surface by a few millimetres so it did not touch it any more in the second row while still being in the proximity of the piezo ring.

The diamond tip on the pickup needle can swing in two directions, up-down and sideways with respect to the surface of a record on a turntable. When the needle tip follows the groove coined into the surface of a stereophonic record, the 2D plane formed by these two degrees of freedom is used to encode the two-channel information of the stereo sound. But for the sake of physical symmetry, the two stereo channels are encoded in a coordinate system rotated by 45° relative to the "up-down-left-right" coordinate system. During the measurement campaign just one channel was recorded because the difference to the surface-perpendicular motion component distilled properly from the two stereo channels would in practice just have been a proportionality factor.

It is the tininess in size and excitation force of the pickup needle and its magnet-coil transducer that makes it suitable for probing the resonator glass wall displacement without influencing it throughout a wide frequency band. The eigenfrequency of the pickup needle is beyond the frequency range of interest which goes up to 22 kHz [5]. One of the few requirements to let the probing system function properly is the right contact force of the needle on the probed surface. Figure O.4 illustrates the construction of the needle bearing which ensured that this force stayed within the suitable range.

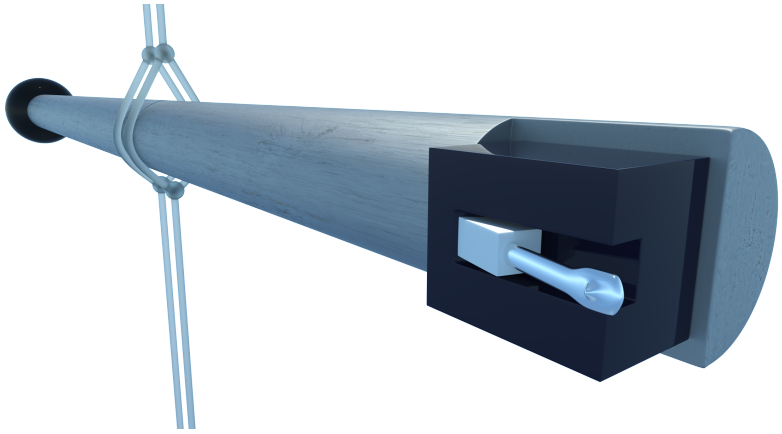


Figure O.4 The bearing for the pickup needle cartridge. The displacement pickup needle (Ortofon® OMB 5E) was mounted on a thin wooden arm and balanced with counterweight. The arm was supported in the middle by two shackles being part of two vertical nylon strings. This created a frictionless bearing with rotational freedom for the wooden arm (the strings form the rotation axis) and a very low torsional spring constant. Pulling at one arm with a newtonmeter allowed to determine the degree of arm deflection corresponding to a suitable contact pressure (between 15 and 35 mN) for the pickup needle. (Artwork: Tudor Pirvu)

The signal line: The small signal from the magnet-coil transducer behind the pickup needle was amplified using the phono input of a Realistic® STA-530 stereo HiFi amplifier. A sketch is shown in figure O.5. Only one of the two stereo channels from the cartridge was amplified and stored via Labview. The aluminium foil shielding around the cartridge-to-coax cable connection area was connected to the amplifier's ground potential. The two leads of the amplifier's speaker output channel in use were connected with two resistors in parallel yielding together 55 Ω. The voltage drop across was the signal taken to the ADC board. The transfer function of the STA-530 amplifier was roughly determined at 5, 10, and 20 kHz

by replacing the needle cartridge in the above-described setup with the HP 33120A function generator. The result is shown in figure O.6. However, the transfer function, in particular the phase shift, of the whole instrumentation chain, including the needle cartridge's magnet-coil transducer, could not easily be determined because of a lack of independently generated absolute wall displacement data.

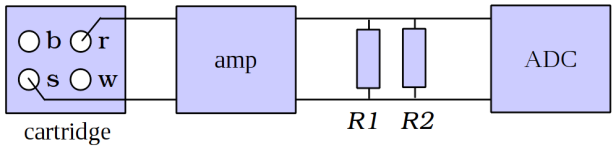


Figure O.5 Schematic of wall displacement measurement signal line.
The pickup needle cartridge has four electric pins. Three of them have colour labels in blue, red and white. The remaining pin is connected to the metal shielding box of the magnet-coil transducer and should thus always be connected to ground. The two stereo channels are blue-white and red-shield. The pattern is recognisable since the channels have an internal resistivity of ≈700 Ω [5]. One channel has been selected, amplified with the Realistic® STA-530 HiFi amplifier, and the voltage drop of its speaker output across two resistors was the recorded signal. The two resistors were of a type with an elevated power dissipation ability. Their resistivities were 110.8 Ω and 109.2 Ω, yielding 54.97 Ω in parallel.

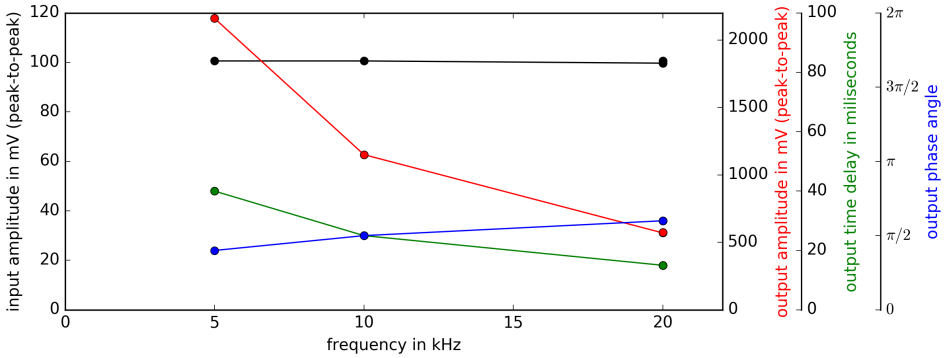


Figure O.6 The transfer function of the Realistic STA-530 amplifier.
It has been tested with a sinusoidal input signal directly from the HP 33120A function generator. The data plotted in black shows the constant voltage amplitude of the function generator output. The red and green data show the amplitude and time shift of the sinusoidal output of the amplifier. Shown in blue is the phase angle calculated from the time delay.

O.1.7 Digitisation and basic analysis of harmonic signals

Analogue-to-digital conversion (ADC)

After converting the physical signals (displacement, pressure, electric current, ...) into voltages they were digitised either with the two GPIB-capable oscilloscopes (before case 200) or with the PC-mounted ADC card (after case 200). The two scopes allow the sampling of four signals in parallel upon one single trigger signal with a sample rate of 2×10^5 samples per second. The ADC card is able to scan a single channel at a rate of $1.24 \times 10^6\,\mathrm{s}^{-1}$ or n channels in parallel at $1/n$ times the rate. When the ADC card was used for analysing the stationary resonator response, the multiple signals have therefore been recorded one at a time exploiting

the function generator as trigger several times in sequence. (Some more details are given in appendix P.1.3.)

Determining the frequency response

The digitised snippets of the oscillating voltage and current signals can be analysed employing Fast Fourier Transform (FFT) routines.[6] Working with the idealising assumption that the resonator under examination exhibits a linear response to the sinusoidal forcing through the voltage $U(t) = U_0 \sin(\omega t)$, each of the response signals itself is by definition a sinusoidal signal of the form $s(t) = s_0 \sin(\omega t + \varphi_s)$. As the two parameters amplitude s_0 and phase φ_s (at the excitation frequency) suffice for a complete specification, it is all that needs to be taken from the FFT result $\hat{s}(\omega) : \mathbb{R}^+ \to \mathbb{C}$. The frequency response of the resonator is obtained automatically by a Labview code (see appendix P.1 for details) which steps through the frequency interval of interest, buffering and analysing a snippet at each step after waiting for the establishment of stationary oscillation and storing only the values s_0 and φ_s.

Deduced electrical quantities

Looking at any transducer at a given working point, i.e. while oscillating harmonically in a stationary state, voltage (U_0, ϕ_U) and current (I_0, ϕ_I), can be interpreted as points in the complex plane,

$$U = \operatorname{Re} U + i \operatorname{Im} U = U_0 \left(\cos(\phi_U) + i \sin(\phi_U)\right) \tag{O.1}$$

$$\text{and} \quad I = \operatorname{Re} I + i \operatorname{Im} I = I_0 \left(\cos(\phi_I) + i \sin(\phi_I)\right), \tag{O.2}$$

and Ohm's law, $U = RI$, turns into $U = ZI$, yielding the definitions of impedance $Z \in \mathbb{C}$ and admittance $Y = 1/Z$. The response functions $Z(f)$ and $Y(f)$ contain essential information about the electromechanical properties of the resonator, as outlined in appendix J.

O.2 Characterising single piezo rings

This appendix chapter is about presenting the characterisation data of transducer-driven resonators. Let's begin with the simplest system: the transducer alone, without the resonator.

In the design by C. West and R. Howlett [505] a cylindrical glass flask is set in motion by a hollow cylinder of radially polarised polycrystalline piezoelectric ceramic. The outer and inner surfaces of such a ceramic hollow cylinder are coated with thin metal layers forming the electrodes of the transducer. An electric potential between the electrodes creates an electric field parallel to the macroscopic polarisation vector of the polycrystalline ceramic. The piezoelectric effect induces a strain in the material as consequence of the $\vec{E}$-field and this leads to a deformation of the hollow cylinder, either a growing radius with a thinning of the wall thickness, or the other way round. The transducer's response to an alternating voltage is the excitation of a radial displacement oscillation.

[6] See appendix P.1.3 for implementation details.

For all resonators examined within this project the transducers were hollow cylinders made of PZT-8[7] with an inner diameter of 65 mm, a height of 25 mm, and a wall thickness of 3 mm. The geometrical tolerance after machining of the sintered parts is given by the manufacturer as 0.13 mm [353].

O.2.1 Experimental setup

Two such piezo rings were examined, one had been salvaged[8] from the broken resonator Nº 7, the other one came new from Channel Industries®. They were hung up, their rotation symmetry axis kept vertical, by suspending them with soft cotton yarn spiralled several times through the hollow cylinder and around a horizontal steel rod 20 cm above it, so the weight of the piezo ring was distributed onto many equidistant support points. Apart from the tolerances of the manufacturing process, the solder joints and the wiring on the electrodes are additional sources of asymmetry of the setup. Only electrical properties were recorded from the piezo rings.

O.2.2 Measurement results

Figure O.7 shows the electrical raw data gained from frequency sweeps on two unloaded transducers. The data sets correspond to the measurements documented as cases 105 and 114. For the sake of clarity no different labelling is introduced in this text besides the numbering of the lab records. The data sets were gained with the same experimental setup, but examining two different PZT rings with two different histories. One case corresponds to a used and salvaged transducer ring, the other to a new one. Their resonance frequencies are around 280 Hertz apart. How much of this offset should be attributed to the difference in history and how much to the difference in production batch cannot be determined from these two samples.

Figure O.8 shows electrical raw data gained in a setup where the transducer was driven directly by the amplifier with the transformer shorted or taken away. This keeps the voltage more stable. As a consequence, around the resonance, both voltage and current are available with a high signal-to-noise ratio. This allows the deduction of a clear phase signal (using the Labview pattern shown in figure P.1). Only around the antiresonance a region remains where one of the two primary signals, the current, exhibits very low amplitudes resulting in a degrading quality of the phase signal. This situation translates into clean admittance circles and less clean impedance circles.

Admittance circle analysis of an unloaded transducer

The finely resolved resonance recording of case 158 can be used for an admittance circle analysis. Figure O.9 shows the raw data, figure O.10 the computed values of

[7] PZT is short for lead-zirconate-titanate $PbZr_xTi_{1-x}O_3$. PZT-8 is a Navy type III material with $x = 0.52$ [263]. See also appendix Q.1.2.

[8] Therefore, it differed from the other one in three aspects: (a) it was part of a different production batch, having been manufactured years before, (b) it had been in use on resonator Nº 7 for a short time, and (c) it had been in a chemical bath dissolving the two-component epoxy glue over several days.

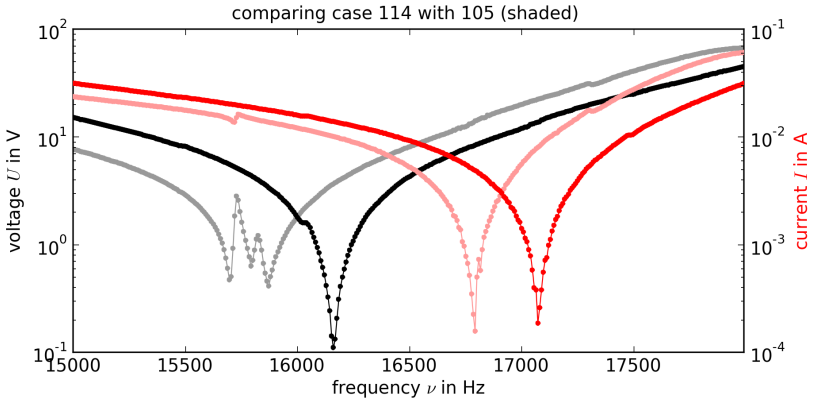


Figure O.7 Raw electrical measurement data from two unloaded transducers
The plot shows the amplitudes of voltage (black) and current (red) signals feeding the PZT-8 hollow cylinders over the frequency axis. The shaded plots correspond to the salvaged transducer (case 105 in the lab logbook) while the fully coloured ones belong to the unused ring from a more recent manufacturing batch (case 114). The measurements were made with the transformer behind the amplifier as depicted in figure O.1. In this setup the resonance and antiresonance can be seen as dips in the voltage and current. One difference between the two transducers is that the frequency pattern is shifted about 280 Hertz, the second difference is the multi-valley structure of the resonance exhibited by the salvaged piezo ring. Whether the differences are the result of the usage history or the manufacturing process cannot be told from these two available samples.

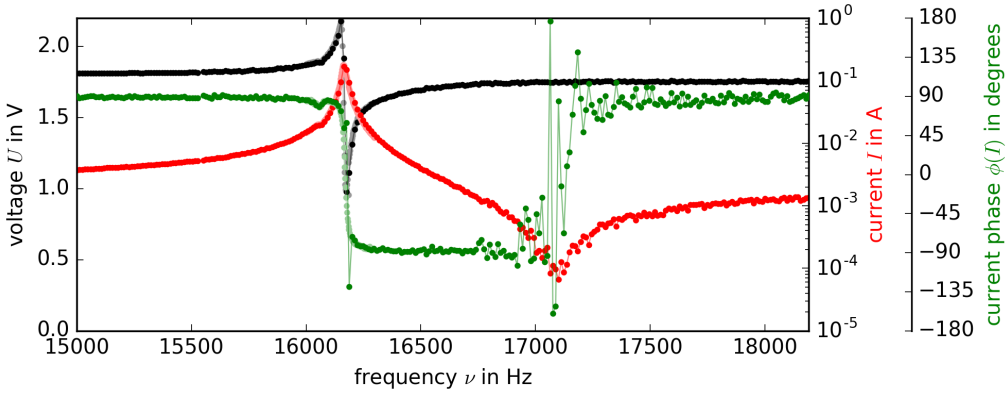


Figure O.8 Raw data from new transducer, resonance and antiresonance
This frequency sweep was recorded with the transformer taken away resulting in an almost stable voltage amplitude. The resonance and antiresonance are now revealed by a maximum and a minimum in the current amplitude. The degradation of the phase signal around the antiresonance is due to the low current amplitude leading to a low signal-to-noise ratio of the current signal. Two frequency sweep records are actually plotted in the diagram: case 157 (full colours) covers the entire plotted domain with a step size of 12 Hz; case 158 (shaded colours) is a close examination of only the resonance with a step size of 2 Hz.

impedance Z and admittance $Y = 1/Z$ as functions of frequency, and figure O.11 shows a plot of the same data in the complex admittance plane. In that latter one, the whole data set appears as grey crosses. The blue dots represent a manually selected subset, and the green ring is a circle numerically fitted to that subset by minimising[9] the sum of square distances of the data points from the circle. The fit parameters were only the x- and y-coordinates of the circle centre, and not the

[9]The utilised search algorithm was Powell's method [361] from the Python library SciPy [223] (version 0.9.0). It can be found there as `scipy.optimize.fmin_powell`.

radius, because the latter was set to equal the x-coordinate in accordance with the simplified transducer model of appendix J. The data subset marked in blue is thought to be more suitable than the entire data set for the circle fitting procedure for two reasons: on the one side the blue set is more evenly balanced around the circle, and on the other side the neglected densely packed data points corresponding to lower values of the conductance G are far away from the resonance on the frequency axis. The fitted circle centre has an x-coordinate of 73.61 mS and a y-coordinate of 296.6 µS. The mean distance of the data points (the blue subset on which the fit is based) from the circle is 668.9 µS

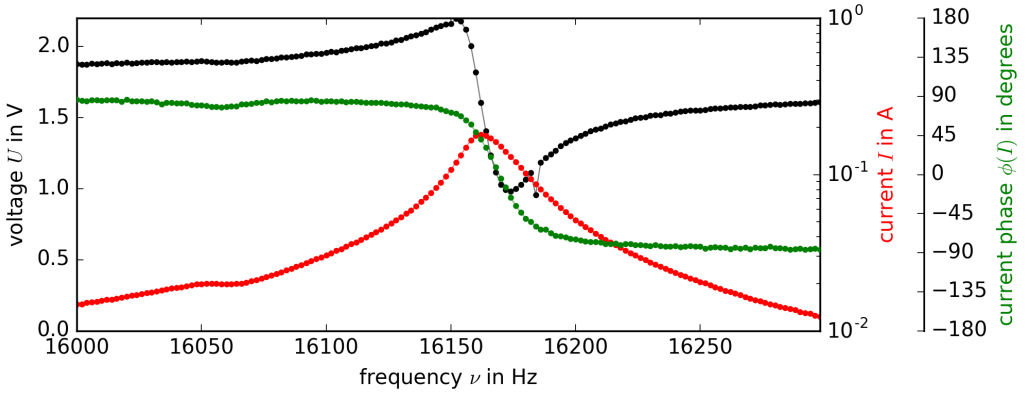


Figure O.9 Raw data of finely resolved resonance from the unloaded new transducer.
(Data of case 158)

According to the second formula of table J.2 the mechanical Q-factor of the transducer can be calculated from the frequencies f_{mB}, f_s, and f_{nB} corresponding to the points of the circle at the top, the right, and the bottom. Making linear interpolations of the discrete data set, these frequencies are 16160.92, 16169.31, and 16 178.50 Hz, respectively, from which follows

$$Q_m = \frac{f_s}{f_{nB} - f_{mB}} = 919.79 \approx 920.$$

The electric Q, which is $Q_e = B_s/G_{\max}$, by contrast, cannot reliably be determined with the present data set because $B_s = 296.6\,\mu\text{S}$, the vertical offset of the circle centre from the real axis, is substantially smaller than the mean distance of the data points to the fitted circle (668.9 µS), so that Q_e appears to be zero through this first approach. One could perhaps say that this implies an upper limit $B_s < 0.7\,\text{mS}$.

Next, the sizes of the equivalent circuit elements can be calculated as

$$R = \frac{1}{G_{\max}} = 6.80\,\Omega, \qquad L = \frac{Q_m R}{\omega_s} = 61.5\,\text{mH}, \qquad C = \frac{1}{Q_m R \omega_s} = 1.58\,\text{nF}.$$

The parallel capacitance C_0 (i. e. the leak path existing for AC current due to electrodes and cables forming a capacitor) cannot be inferred from $C_0 \approx B_s/\omega_s$ and must instead be calculated as

$$C_0 = \frac{f_r^2}{f_a^2 - f_r^2} C. \tag{O.3}$$

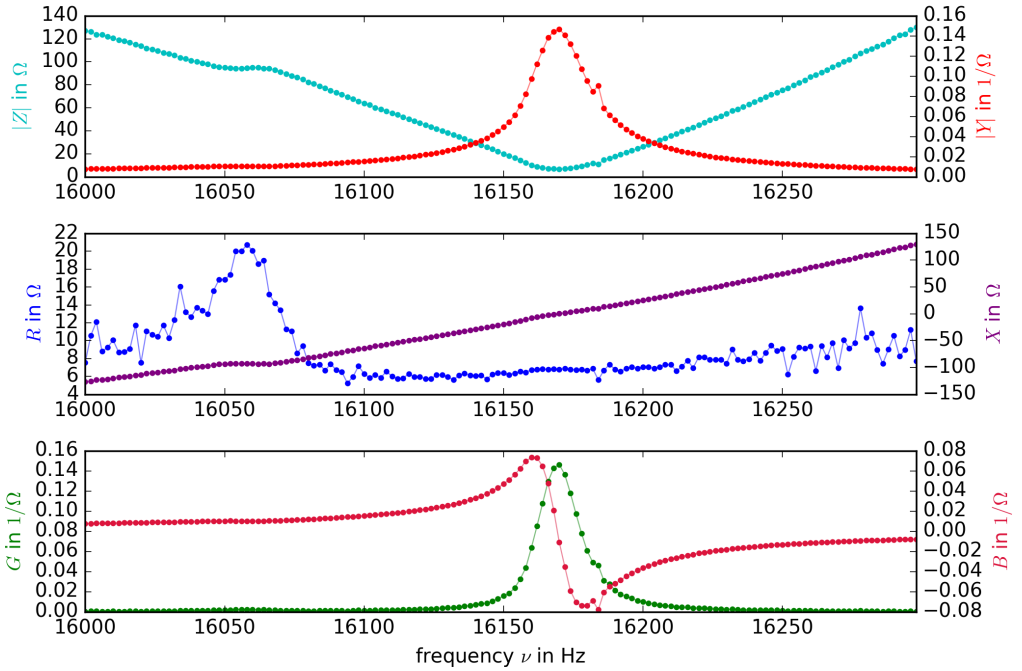


Figure O.10 Impedance and admittance of an unloaded transducer.
This plot shows the real and imaginary parts of admittance $Y = G + iB$ and impedance $Z = R + iX$ as well as their magnitudes $|Y|$ and $|Z|$.

The antiresonance frequency f_a is needed here, and it must be read from the data set 157 plotted in figure O.8, where the quality of the phase signal is unfortunately very low above 17 kHz due to the current amplitude minimum at the antiresonance.[10] From that data the antiresonance frequency f_a was determined to be $f_a = (17120 \pm 30)\,\mathrm{Hz}$ with the justification presented in appendix P.3.1. Together with $f_r = 16\,169.3\,\mathrm{Hz}$ this yields

$$C_0 = \frac{f_r^2}{f_a^2 - f_r^2} C = rC = 13.01\,\mathrm{nF}.$$

With the uncertainty on f_a being 30 Hz and neglecting the uncertainty on the other variables against it, the uncertainty on C_0 turns out to be 0.4 nF. Hereby, a capacitance ratio of $r = C_0/C = Q_e Q_m = f_r^2/(f_a^2 - f_r^2) = 8.26$ was used. Since r is at the same time the Q-product, an electric Q-factor of $Q_e = 8.98 \times 10^{-3} \pm 2.9 \times 10^{-4}$ can also be computed. Now, B_s can be calculated again using the formula $B_s = \omega_s C_0$,

[10] At the same time it becomes clear that the data quality degradation around the antiresonance poses no severe problem for the transducer characterisation in this case, where we are dealing with a high-Q system. The quantities L, R, and C can all be gained solely with the clear admittance circle data stemming from the resonance, and with that the physical properties of the transducer itself are fully determined. Only for the determination of C_0 which quantifies the AC leak path through electrodes and cables the data from the antiresonance is necessary, and it contributes to the formula only through a term of the form $f_a^2 - f_r^2 = (f_a + f_r)(f_a - f_r)$. In a high-$Q$ system where $f_{nX} - f_{mX} << f_a - f_r$, i. e. where the phase transition region is much smaller than $f_a - f_r$, it is in this case not so important to have the antiresonance well resolved, it is more important to know where it is. See also figure P.6 which addresses the same point.

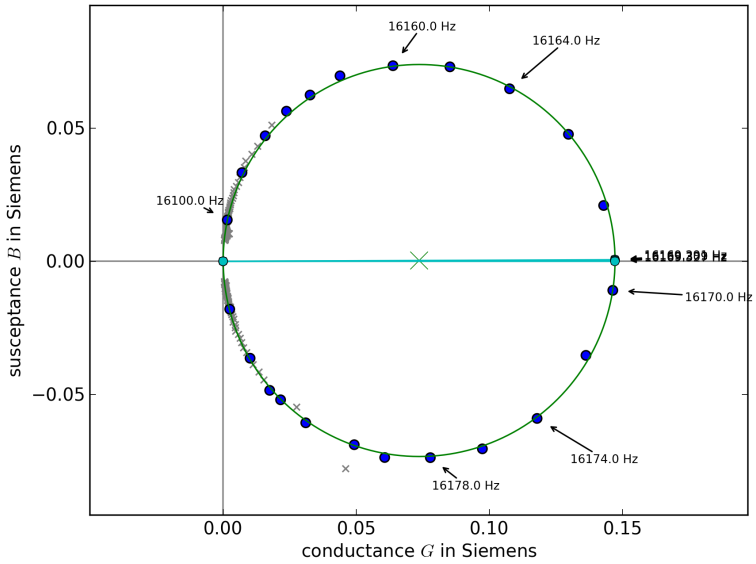


Figure O.11 Admittance circle of the unloaded new transducer.
This diagram shows the admittance data in the complex plane $Y = G + iB$ where the excitation frequency is a parameter. The grey crosses reflect the whole data set (case 158, as in figure O.10) while the blue dots represent a manual selection deemed most suitable for numerically fitting a circle. The best fitting circle is shown in green. The boundary condition of the fit procedure was that the imaginary axis had to be a tangent of the circle. This means that whereas a circle in a plane is normally defined by three parameters (coordinates of the centre and radius), the admittance circle fitting procedure involves only two free parameters.

and one can obtain a value of $B_s = (1.32 \pm 0.043)\,\mathrm{mS}$, whereby the uncertainty still stems from the uncertainty on f_a deemed to dominate the others. The contradiction of that value and the estimate $B_s < 0.7\,\mathrm{mS}$ stated above may be due to the many idealising assumptions at the basis of table J.2.

The electromechanical coupling factor k and the figure of merit M can be determined via

$$k = \sqrt{1 - \frac{f_s^2}{f_p^2}} = 0.33 \pm 0.005, \qquad M = \frac{k^2 Q_m}{1 - k^2} = 111.34 \pm 4,$$

whereby for f_p it was assumed $f_p \approx f_a$.

Lastly, the motional capacitance constant Γ can be specified for this PZT hollow cylinder with a thickness of $d = 3\,\mathrm{mm}$, a height of $h = 25\,\mathrm{mm}$, and an inner radius of $r_i = 32.5\,\mathrm{mm}$:

$$\Gamma = \frac{Cd}{A} = \frac{Cd}{2\pi r_m h} = 8.88 \times 10^{-10}\,\mathrm{F\,m^{-1}},$$

where r_m is the radius of the middle of the cylinder wall.

O.3 Characterising resonator no. 8

Resonators № 7 & 8 were the first SF resonators built at RPI according to the design with H-shaped cross section developed by Cancelos [69]. The new design

reflected (a) ab initio thoughts on the desired sound field, (b) a simulation-aided approach to understanding the connection of the sound field with the structural vibration mode shape, and (c) the wish to make the manufacturing and assembly process more reproducible. Unfortunately, both resonators disappointed the hopes to furnish better-performing SF resonators. Resonator 7, made of a single glass piece, suffered a fracture in the glass at a point of high wall curvature early on during experimentation. № 8 turned out to be a system with a far too low Q-factor, which is documented below.

O.3.1 Cavitation experiments with resonator no. 8

The SF experiment is feasible only if one can maintain a decent rate of cavitation bursts inside the acoustic resonator over hours while a cool temperature of $\sim 0\,^{\circ}\mathrm{C}$ is kept stable. This can be tested with normal instead of deuterated acetone and with the PuBe neutron source instead of the pulsed neutron beam for bubble nucleation. Such cavitation tests yielded, however, very disappointing results with resonator № 8.

The tests were conducted at the Gaerttner Laboratory at RPI. With the resonator placed in a freezer, under video supervision for visual cavitation detection and in proximity of a fan for cooling via streaming air, the following sequence of events and observations could be conducted several times. Before filling the acetone into the resonator, it has to be degassed in a glass flask, where a large free surface can be exposed to the gas phase, by agitation through a magnetic stirrer and while maintaining the vapour pressure above the surface with the help of a vacuum pump. Once transferred into the resonator the degassing process of the liquid needs completion through acoustic agitation. Setting the resonator in vibration while maintaining vapour pressure in the connected vacuum system leads to bubbling like in soda water throughout the resonator volume even at low driving voltage. With the progress of the degassing process the resonator slowly turns into a neutron detector, requiring greater driving voltages and the presence of the neutron source for triggering new bubble clouds once a bubble population dies out. It becomes improbable to see soda-like sparkling throughout the chamber and the phenomenon visible then is a dancing jet of bubbles originating at the sound pressure antinode and pointing outwards in often wildly changing directions. The jet becomes intermittent and reappears only in the presence of the neutron source. The jet lifetimes shorten. When their lifetime becomes so short that they become mere bursts, then the working regime for SF experiments is reached. Now, according to Taleyarkhan's SF protocol, the voltage amplitude representing the cavitation threshold has to be sought and from there the driving voltage should be doubled.

With resonator 8, cavitation led to heating and subsequent boiling and required the full power of the driving electronics so that a doubling of the voltage was impossible. Already during the bubble jet phase the three thermocouples placed on the aluminium end plates and on the glass wall measured increasing temperatures (e. g. from 6 to 14 °C in 5 minutes) indicating that the heat removal via fanned air (with temperatures between -5 and 0 °C) was largely insufficient. Only very short phases of the cavitation burst regime could be observed (e. g. 15 seconds) before boiling be-

gan. Driving voltage amplitudes between 600 and 900 Volt were necessary to reach the cavitation burst regime. Beyond 950 V the system showed large nonlinearities (overtones in the scope signals) and the "peak" warning light on the amplifier went on. After each short cavitation trial thermocouple temperatures peaked at 20 to 35 °C. Several minutes of cooldown time without acoustic drive were then needed. Repeated cavitation trials were conducted with letting the chamber cool down to about 10 °C in between. Frequency sweeps at lowered power were used to determine the frequencies of maximum microphone signal amplitude (found in between 19.4 and 19.7 kHz) and maximum microphone gain (18.8 to 19.2 kHz).[11] Both, frequencies of maximum gain as well as frequencies of maximum amplitude were used to test the cavitation behaviour. The cavitation thresholds were always between 70 and 90 % of the linear range of the amplifier power with a tendency to be slightly lower in the maximum gain cases. Boiling bubbles were nucleated at the top and bottom flanges. Whether a small fraction was also created behind the transducer or just grew bigger while passing it was difficult to tell.

In an attempt to get a hint whether the increased use of the rather soft metal aluminium is the reason for the high heat generation rates, the bottom piston was replaced by a new construction involving a glass piston of similar shape being clamped between the rubber seal and an aluminium ring. The design is depicted in figure O.12. However, repeated cavitation tests brought no improvement: almost only bubble jets and very little cavitation could be observed, and after minutes the acetone boiled at temperatures above 10 °C. Boiling bubbles also came from the bottom of the chamber. This means that either the heat generation takes place in the bending glass at similar rates as in aluminium (and is not seen at the glass side walls because they have smoother surfaces) or it may be due to frictioning between assembly parts which are bolted and clamped against each other.

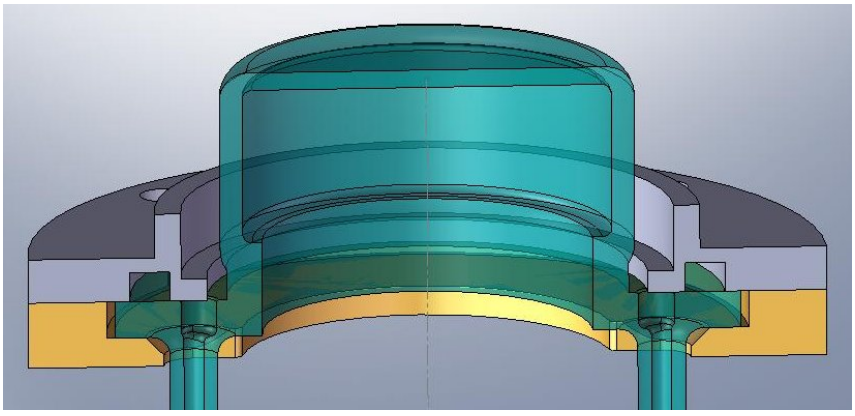


Figure O.12 Bottom flange modification of resonator Nº 8.
This alternative bottom head piece was constructed and tested seeking more information on what determined the energy dissipation of resonator 8. It is a glass piece clamped between an aluminium ring (beige) and the rubber O-ring. The glass part consists of pieces cut out from glass plates and tubes fused together with glassblowing techniques. Grinding and polishing was necessary for the surface pressed against the rubber seal (to be placed into the groove of the aluminium counterpart coloured in grey).

[11] Figure O.18 shows how these frequencies shift within these ranges depending on the temperature.

As a consequence, this resonator design is not suitable for the SF experiment in the current setup. One could ask whether it could be made suitable in connection with driving and cooling systems of increased power. In any case, material limits like the maximally allowable[12] potential gradient in the piezoelectric material should be taken into account. The other principal problem with a design relying on bolted and clamped part connections is that simulations (e. g. FEM) are difficult and unreliable.

O.3.2 Electrical properties

The unmodified version

The measurements with resonator Nº 8 were all conducted with the transformer in between amplifier and transducer, in a setup where the voltage amplitude varies greatly between different working points. The amplitude minima of both voltage and current do not hinder the admittance and impedance circle analysis in this case because the low Q-factor prevents these dips from reaching down extremely low. Figures P.8 and P.9 (page 404f) show the electrical raw data and the deduced impedance and admittance of resonator Nº 8. The gained admittance and impedance circles are presented in figures O.13 and O.14.

Analysing the admittance circle, the resulting mechanical Q-factor turns out to be

$$Q_m = \frac{f_s}{f_{nB} - f_{mB}} = 59.55 \approx 60,$$

while the electrical one is

$$Q_e = \frac{B_s}{G_{\max}} \approx 0.34.$$

In comparison with the unloaded transducer analysed in the preceding section, it can be seen that the drastically lowered mechanical Q reflects the resonance widening, the fact that a much larger frequency interval is needed for the phase angle transition which can also be seen by the much narrower data point spacing along the Y and X circles.

The equivalent circuit elements R, L, and C evaluate to

$$R = \frac{1}{G_{\max}} = 405\,\Omega, \qquad L = \frac{Q_m R}{\omega_s} = 212\,\mathrm{mH}, \qquad C = \frac{1}{Q_m R \omega_s} = 0.364\,\mathrm{nF}.$$

The increased dissipation of the transducer load can here be seen by the much larger R in comparison with the unloaded case, while the larger L reflects the added inertia of the structural masses in oscillatory motion. Since the capacity C functions mathematically as the reciprocal of the spring constant of the oscillator model, it has to decrease (i. e. the system has to become stiffer) in order to account for the not much changed resonance frequency while the mass more than tripled.

The parallel capacitance C_0 can now be computed in two ways,

$$C_0 \approx \frac{B_s}{\omega_s} = 7.35\,\mathrm{nF}, \qquad C_0 = \frac{f_r^2}{f_a^2 - f_r^2} C = 18.1\,\mathrm{nF}.$$

[12] [353] gives a limit of $8\,\mathrm{kV\,cm^{-1}}$. A voltage drop of $\approx 900\,\mathrm{V}$ across the 3 mm thick transducer was the maximal voltage applied to resonator 8. In the experimental setup this limit was given by the linear range of the amplifier.

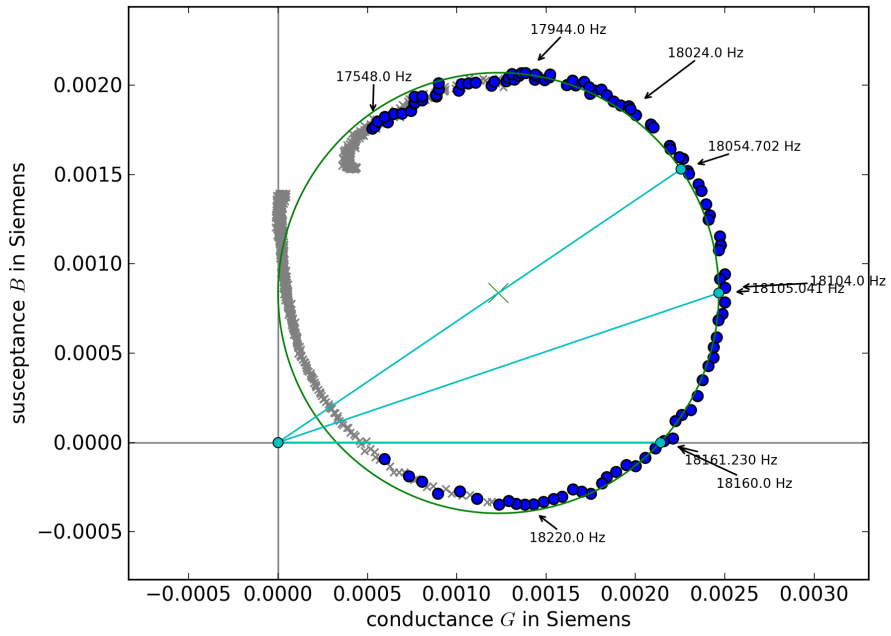


Figure O.13 Admittance circle (data and fit) of resonator Nº 8.
Again, the grey crosses show the whole data set (case 25) gained by the frequency sweep and the blue circles correspond to the subset having been used for fitting the green circle. The selection has not been conducted completely manually, but just by deliberately specifying ignoring ratios (e. g. for all points with $G/G_{\max} \in [0.2, 0.3]$ five out of six points were ignored, for $G/G_{\max} \in [0.3, 0.5]$ two out of three, and for $G/G_{\max} > 0.5$ none). The green circle is determined by $G_{\max} = 2.467\,\mathrm{mS}$ and $B_s = 0.8366\,\mathrm{mS}$. The characteristic frequencies not indicated in the plot are $f_{mB} = 17.9217\,\mathrm{kHz}$ and $f_{nB} = 18.2255\,\mathrm{kHz}$.

where $f_a = 18.343\,\mathrm{kHz}$, $f_r = 18.161\,\mathrm{kHz}$, and $f_s = 18.105\,\mathrm{kHz}$ are the antiresonance, resonance, and the series resonance frequencies which can be read from diagrams O.13 and O.14. The Q product or capacitance ratio $Q_e Q_m = r = C_0/C$ also evaluates differently, depending on the formula:

$$Q_e Q_m \approx 20.2, \qquad \text{whereas} \quad r = \frac{C_0}{C} = \frac{f_r^2}{f_a^2 - f_r^2} \approx 49.8.$$

The electromechanical coupling factor k and the figure of merit M evaluate to

$$k = \sqrt{1 - \frac{f_s^2}{f_p^2}} = 0.22 \qquad \text{and} \quad M = \frac{k^2 Q_m}{1 - k^2} = 1.8,$$

and the motional capacitance constant Γ to

$$\Gamma = \frac{Cd}{A} = \frac{Cd}{2\pi r_m h} = 2.04 \times 10^{-10}\,\frac{\mathrm{F}}{\mathrm{m}}.$$

The modified version with a glass bottom plate

The analysis of the electrical data (figure P.10) shows that the introduction of the clamped glass bottom plate increases the losses, as can be seen by the lowered Q_m

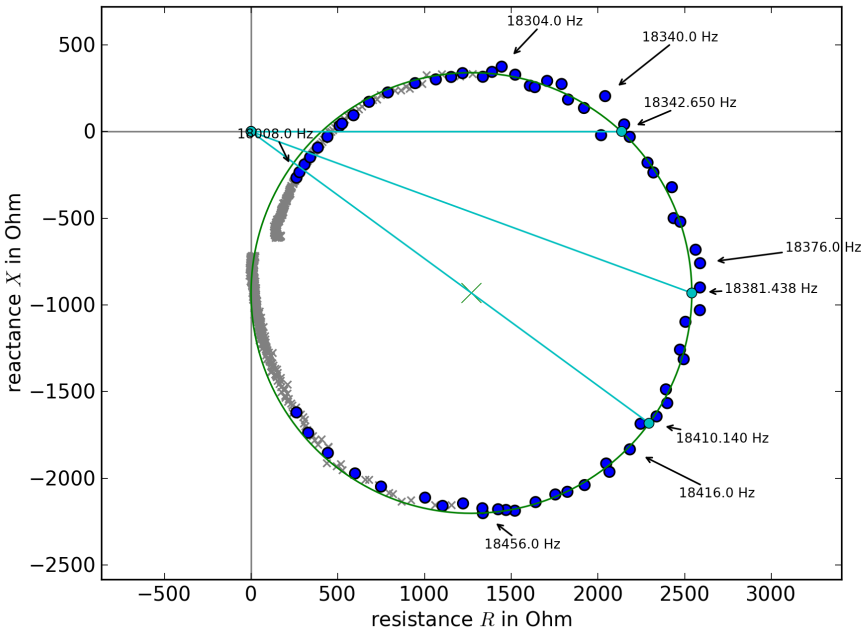


Figure O.14 Impedance circle (data and fit) of resonator № 8.
The quantities determining the fit circle are $R_{\max} = 2.542\,\mathrm{k\Omega}$ and $X_0 := X_p = -931\,\Omega$, the reactance at f_p. (Data of case 25)

and increased Q_e in table O.2. Damping mechanisms suppress the dip in the phase signal shrinking the Y- and Z-circles and together with the increased C_0 observed with the modified setup this pushes their centres sufficiently far away from the real axis so there are no intersections any more. This can be seen in figure O.15.

Table O.2 Electrical properties in comparison: two setups of resonator № 8.

quantity	formula	alu piston	glass piston
Q_m	$\frac{f_s}{f_{nB}-f_{mB}}$	60	42
Q_e	$\frac{B_s}{G_{\max}}$	0.34	0.64
R	$\frac{1}{G_{\max}}$	405 Ω	490 Ω
L	$\frac{Q_m R}{\omega_s}$	212 mH	174 mH
C	$\frac{1}{Q_m R \omega_s}$	0.364 nF	0.408 nF
C_0	$\frac{B_s}{\omega_s}$	7.35 nF	11.02 nF
k	$\sqrt{1-\frac{f_s^2}{f_p^2}}$	0.22	0.19
M	$\frac{k^2 Q_m}{1-k^2}$	1.8	1.4
Γ	$\frac{Cd}{A}$	$2.04 \times 10^{-10}\ \frac{\mathrm{F}}{\mathrm{m}}$	$2.29 \times 10^{-10}\ \frac{\mathrm{F}}{\mathrm{m}}$

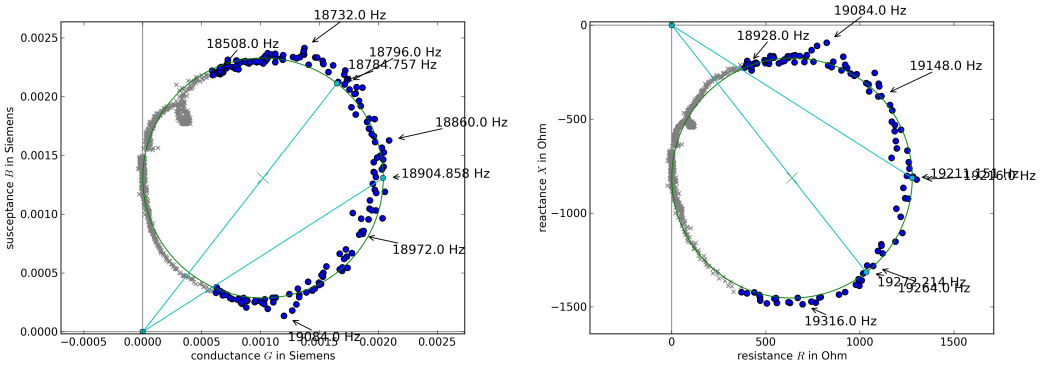


Figure O.15 Admittance and impedance circles of modified resonator Nº 8.
The fitted Y-circle is given by $G_{\max} = 2.042\,\mathrm{mS}$ and $B_s = 1.309\,\mathrm{mS}$. The characteristic frequencies not indicated in the plot are $f_{mB} = 18.6599\,\mathrm{kHz}$ and $f_{nB} = 19.1091\,\mathrm{kHz}$. The quantities determining the Z-circle are $R_{\max} = 1.277\,\mathrm{k\Omega}$ and $X_0 := X_p = -813.2\,\Omega$, the reactance at f_p. The additional structures between 17.7 and 19 kHz translate into the little loop visible in both plots between ten and eleven o'clock. The selection (blue) of the data (case 47) to be fitted was done simply by threshold, whereby $G/G_{\max}$ and $R/R_{\max}$ had to be greater than 0.3.

O.3.3 Acoustic properties

Resonator 8 is the only one allowing the use of the hydrophone while the shape of the liquid volume remains unchanged because the top head of the chamber is made of aluminium and a special version of it with an additional outlet for holding the hydrophone could be easily manufactured. The outlet is in the centre of the top plate and the hydrophone can be fixed and sealed tight in it by a rubber O-ring squeezed by a nut. The fixation allows variable vertical positions of the thin cylindrical hydrophone so it can probe the sound pressure along the resonator's central axis from the top to the bottom piston. A corresponding series of frequency recordings is presented in figure O.16 where it is visualised as a 2D colour map. Two versions of the pressure map are presented, one showing the absolute sound pressure inside the resonator along the central axis of the cylindrical geometry over the frequency and the other one the sound pressure gain. The latter one should give the map that would be observable if it was possible to keep the driving voltage constant. There are three noteworthy aspects: (a) the structures in both 2D maps are not perfectly upright but leaning slightly to the left, (b) the top and bottom halves of the vertical mode shapes are not symmetric, and (c) the gain map reveals a split into three sub-peaks.

The explanation why the vertical structures are leaning to the left is that each peak is systematically shifted a small bit towards higher frequencies when the hydrophone is shifted downwards, because the presence of the hydrophone made of steel increases the overall stiffness of the resonator filling as it replaces the relatively soft liquid. In an alternative thought one can say, assuming the deformation of the hydrophone itself as negligible, that as the radial distance from the central axis to the glass wall is reduced by the hydrophone radius, the standing wave covering that distance has to reduce its wavelength and increase its frequency. Since the structures are left-leaning, it makes a difference, if one wants to plot the vertical pressure profile, whether a vertical cut through the map is taken or whether the profile is

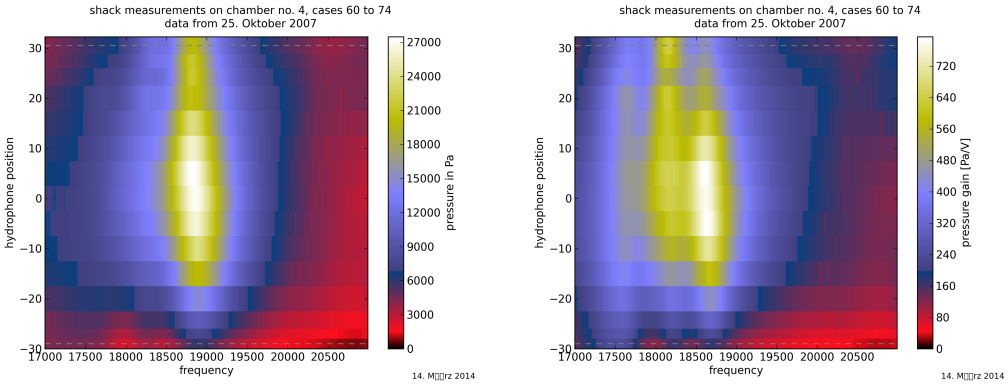


Figure O.16 Sound pressure maps of resonator Nº 8.
These plots represent sound pressure mappings of the resonance near 19 kHz (which is assumed to be the fundamental resonance, where there is one single pressure antinode horizontally and vertically across the main liquid volume). The plot on the left shows the measured sound pressure while the one to the right depicts the pressure gain i. e. the sound pressure gained per unit of driving voltage amplitude supplied to the transducer ring. The hydrophone was moved down in steps of 5 mm, starting at level with the upper piston surface and approaching the lower piston surface to a distance of 1.5 mm. At each step a frequency sweep was recorded (cases 60-74) corresponding to a horizontal line in the plots. During this recording the glass composite bottom piston of figure O.12 had been in place, the setup of figure O.1 was employed, and the signal processing followed the scheme of figure P.1.

composed of local maxima. Such profiles of both sorts of sound pressure and pressure gain are given in figure O.17, and it can be seen that the difference between the vertical cuts and the chains of local maxima is rather small in this case where the peaks are wide.

These profiles visualise the second aspect mentioned above, the asymmetry of the mode shapes even better. This asymmetry is the reflection of the asymmetry of the physical setup, and that has two contributions. For one, the composite glass bottom piston had been already installed, and on the other hand the top piston with the added masses of the hydrophone port and the probe itself surely cannot follow the motion of the liquid as easily as a bare thin plate. In the setup with the two symmetric end plates without hydrophone port installed, it is assumed that the top halves of the pressure profiles take the same shape as the lower halves.

The third mentioned aspect visible in the pressure gain map is the split into multiple peaks. It was observed in microphone gain plots as well. By the microphone records all four setup versions (with hydrophone and the special top head in place or without, with the glass bottom head in place or without) can be compared. But as cases of clean single-peak shapes occurred as well (e. g. figure P.8, page 404), and as no clear correlation with the setup versions could be made out, and finally because no FEM simulation has been attempted of this design with its large frictioning flange connections, no satisfying explanation can be given at this point. (See also figures P.11 and P.12 and caption texts.)

Nevertheless, resonator Nº 8 can still be considered under the viewpoint of the more crucial question whether the resonator functions as intended, i. e. whether the vessel is able to allow the liquid volume to oscillate in its fundamental eigenmode without distorting it. This can be answered largely with “yes”. The pressure shows one single antinode in the centre and it decays sharply towards those volume boundaries formed by piston plates with no hydrophone port. The fact that the hy-

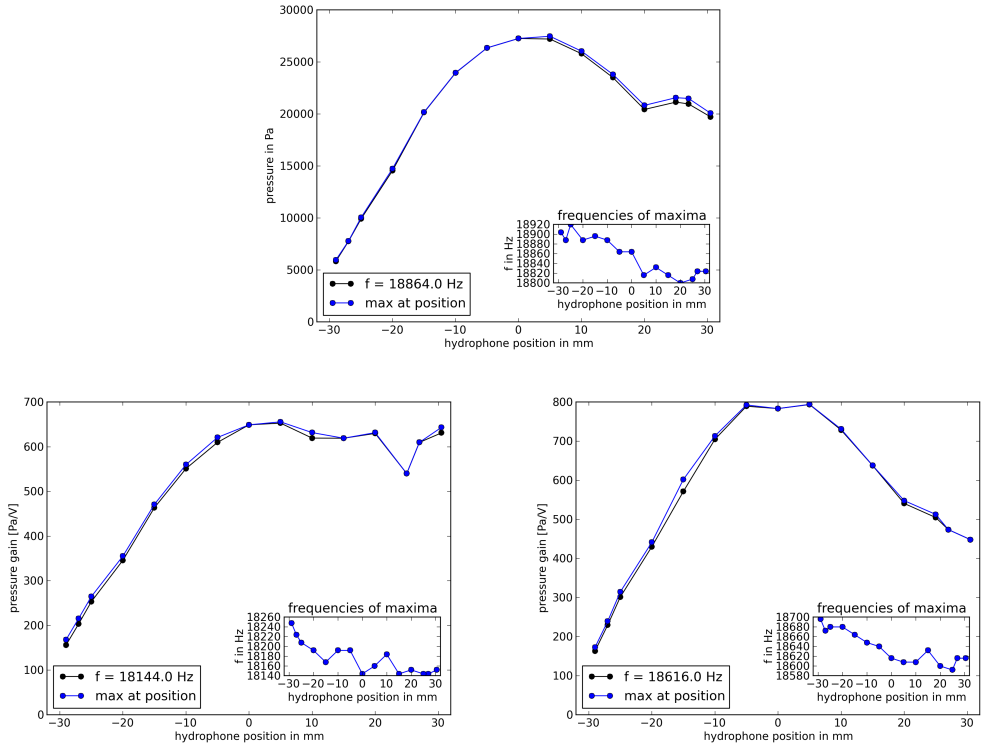


Figure O.17 Sound pressure profiles of resonator Nº 8.
The plot on top represents the vertical profile of the sound pressure map (figure O.16 left) and the ones in the lower row show the profiles of the strongest two of the three peaks visible in the pressure gain map (figure O.16 right). In each case the vertical cut through the pressure map is shown in black and the slanted cut consisting of local maxima is added in blue. The insets show the frequencies of the local maxima and they are able to visualise very well how the lowering of the hydrophone into the liquid volume tunes the resonance frequencies upwards. The pressure response in the centre of the chamber is 1156 $\frac{\text{Pa}}{\text{V}}$ at 18 864 Hz (max pressure) and 1567 $\frac{\text{Pa}}{\text{V}}$ at 18 616 Hz (max gain).

drophone replacing the liquid shifts the frequency upwards is also an indirect proof that the resonance is determined by the liquid rather than the resonator structure. Lastly, the temperature-dependence of the resonance frequency is a fact supporting the identification of the examined resonance with the fundamental acoustic mode. In the raw data presented in figure P.12 in appendix P.5 an exemplary pair of measurements reveals a frequency shift of 500 Hz corresponding to a temperature shift of 16 K. Figure O.18 combines microphone peak frequencies from many characterisation measurements taken with resonator 8 in various setups over a time period of months. That the correlation exhibited by the scattered data stays constant is in agreement with the understanding that this mirrors the connection between temperature and speed of sound in the liquid.

The acoustic signals also allow the determination of Q-factors from peak widths by the formula in table J.2. The Q-factors are of importance when discussing damping and energy dissipation. They are presented in table O.3 and are very low, hinting towards strong damping mechanisms effective in resonator 8. There are several values far below 50. This has to be compared to $Q = 98$ deduced from the hydrophone gain measured on the larger brother of this resonator which had been examined by

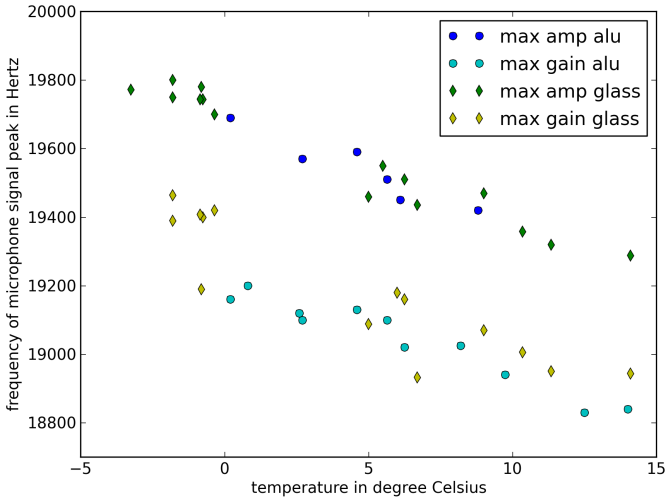


Figure O.18 Resonator № 8 temperature dependence B.
The data cloud shows the dependence of the resonance frequency on the temperature by displaying peak frequencies of the microphone data. The introduction of the glass bottom piston did not change the situation and proves that the resonance is a resonance of fluid motion and as such determined by the liquid's geometry and speed of sound. The speed of sound in acetone is 1174 $\frac{\mathrm{m}}{\mathrm{s}}$ at 25 °C and temperature variations lead to a change by $-4.5\,\frac{\mathrm{m}}{\mathrm{s\,K}}$. Thus, a temperature reduction by 20 °C increases the speed of sound by more than 7 %. The frequency increase of the plotted data is less with only 2.5 % which may be due to the fact that the structural parts play a substantial role in determining the resonances.

Cancelos [69] (reprinted here in fig. I.8, p. 298). She described heat generation as an observable issue, but easily solvable by fanning cool air towards the resonator. During the cavitation trials with resonator 8 this method has proven insufficient. The comparison of dissipation powers (see figure O.19) deduced from the electrical signals can furnish more clarifying context. While Cancelos used the larger resonator at 6 W power for long-term cavitation in water where the cavitation threshold is low, it is a power on the order of 100 W (at driving voltages above 600 to 800 V) necessary for breaching the threshold in cooled acetone with the small resonator version. If one assumes that the two most important energy dissipation mechanisms of this resonator design are (a) frictioning between the tightly clamped metal parts of the flanges and (b) that internal damping (see appendix K.5.8, p. 334) in aluminium is higher than in glass or the transducer ceramic, then two aspects can be explained: On the one hand, the slightly better Q of Cancelos' larger resonator could be due to the smaller proportion of its aluminium parts. On the other hand, the much greater Q-factor difference with respect to the other SF resonators manufactured at RPI can be a consequence not only of a difference of proportion but also of the function of aluminium parts. Resonator 8 is the only SF resonator where aluminium is in contact with the liquid filling. The piston front plates serve as the upper and lower boundaries of the cylindrical main liquid volume. Here, large displacement and small pressure amplitude boundary conditions are required. By contrast, in a resonator design such as resonator 5 it is possible that the upper rim of the main glass wall happens to be a displacement node. It would allow the aluminium flange parts, coupled to the glass wall rim by silicone, to be in a state of small vibration

amplitude. Unfortunately, these thoughts are speculative. It is exceedingly difficult to account for the frictioning bolted connections of flange parts in the FE model. Therefore, these issues could not be explored by simulations.

Table O.3 Resonator № 8: Q-factors from acoustic signals.
The three lacking Q-factors for case 67 are due to multiple overlapping peaks. The exceptionally low peaks of cases 12 & 25 may be due to a seeming single peak being the sum of several ones.

case	setup	T[°C]	hydro	hydro gain	mic	mic gain
12	sym. alu pistons w. hydro	r. t.	15.0	51.8	17.7	52.3
25	sym. alu pistons	r. t.	-	-	19.3	65.7
47	glass bottom piston	6.7	-	-	46.9	32.9
55	glass bottom piston	-1.8	-	-	49.4	50.8
56	glass bottom piston	14.1	-	-	52.9	92.6
67	glass b. piston w. hydro	22.0	30.9	-	-	-

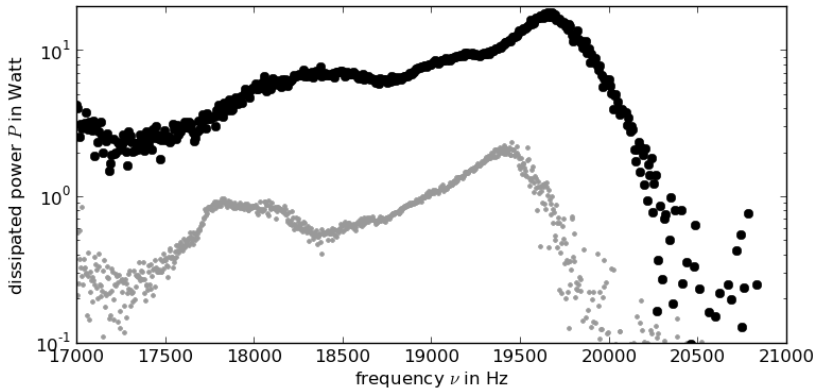


Figure O.19 Resonator № 8: power dissipation at elevated driving voltage.
This plot shows a comparison of dissipated power for cases 47 & 50. The raw data can be seen in figures P.10 & P.11. While the peak driving voltage increased by a factor of 2.1 from 159 to 333 V, the dissipated power grew by a factor of 7.7 from 2.3 to 18.1 W. Another doubling of the driving voltage is needed to breach the cavitation threshold, so extrapolation would imply ~ 150 W heat generation which is similar to a light bulb and would also explain the temperature rises observed during cavitation trials in spite of the fanning of cool air.

Assumed reasons for the high damping rate exhibited by resonator № 8

The reasons for the elevated degree of damping of resonator № 8 may be the choice of materials, a disadvantageous geometry, or the way of connecting assembly parts. The aluminium flanges are assumed to play the central role. Bells producing slowly decaying sounds are usually made of one single piece of cast metal, and everyday experience teaches that bolted assemblies are not able to produce nice bell sounds. Frictioning between the metal surfaces of the flange parts is assumed to contribute to the energy dissipation. A contradiction seems to lie in the fact that resonator 8 was not the first resonator built at RPI with aluminium flanges, but the first one with a particularly low Q-factor, the values ranging from 15 and 60. The seeming contradiction can be explained on the basis of two comparisons: Firstly, the much

larger flange-equipped resonator of the project of Cancelos [69] (depicted in figure I.8) was made for cavitation in water which requires only relatively low sound pressure amplitudes and driving voltages. Secondly, the flange-equipped SF resonators made in the West-Howlett geometry tradition at RPI still have upper pistons made of glass, the aluminium flange consists only of two ring components while the rest of the resonators remained made of glass.

The aluminium parts of resonator 8 differ in function and proportion from their counterparts in earlier SF resonator versions. There is a difference in function because thin aluminium plates are in direct contact with the main liquid and have to fulfill the function of being a boundary condition of large displacement and low pressure amplitude. Then there is the difference in proportion because the aluminium flanges are dominating the design of resonator 8 by their sheer size, whereas for resonators like Nº 5 the flange is a smaller addition. Additionally, it has to be noted that the characterisation setup of resonator 8 always involved two flanges tightly bolted together, whereas resonator 5 was examined either open or with a loosely connected top flange not in contact with the liquid and with the only purpose of holding the hydrophone in its central nozzle.

O.4 Characterising resonator no. 5

After the design of resonator Nº 8 has turned out to result in such strong damping, one of the older resonators was sought for a close examination and comparison. Because of the planned FEM simulations, the even more important purpose was to build a data set for benchmarking the simulations against. Resonator Nº 5 was chosen as the one manufactured and assembled most symmetrically and cleanly, featuring silicone beads with the simplest geometry and of the smallest excess volume. A characterisation signal of primary importance for simulation benchmarking is the output of the sound pressure probe. In order to have the hydrophone in a well-aligned position on the central axis, the specialised top piston with the hydrophone nozzle from resonator 8 was mounted on top of resonator 5. Ideally, the benchmarking data should have been recorded without any top head in mechanical contact with the rest of the chamber because bolted or clamped frictioning contact surfaces are exceedingly difficult to be simulated reliably. But if the acetone filling of the resonator is not kept in a degassed state, only very low driving amplitudes are possible without cavitation occurring on the brushed steel surfaces of the hydrophone whenever one of this resonator's much sharper resonance peaks is met. The aluminium top head with the sealed hydrophone outlet, however, allows to keep the vapour pressure of acetone. For the sake of being able to afford the long times of recording many finely resolved frequency sweeps at decent driving amplitudes consecutively, it had therefore been decided to use the aluminium top, but with tightening the flange bolts as weakly as possible, just enough so degassing with underpressure from a handpump was possible after initially pressing down the top head onto the rubber ring by hand, aiming at the lowest possible degree of mechanical coupling between the main cylinder and the top. All the characterisation measurements were conducted at room temperature between 19 and 22 °C. The vertical positions of both, the hydrophone and the displacement pickup needle, are all given in centimetres

measured from the top rim of the PZT transducer. The same scale applies for the acetone filling level. A filling level of 8 cm was chosen because it leads to a height of the main liquid volume, measured from the lower piston to the free surface, of 11.5 cm, which matches the range of the pressure map presented by Saglime [390] under the condition that the top hydrophone position is coincident with the surface, and that at the bottom position it almost touches the lower piston.

O.4.1 Electrical properties

Admittance circles of very different sizes were recorded with resonator 5. It has already been mentioned that the hydrophone position influences the vibration behaviour. Excluding one source of variation, only measurements corresponding to one single hydrophone position were selected for presentation. Its default position during many measurements was 3.5 cm, which is near the position where the sound pressure maximum can be found; hence, this is the case where the largest number of data sets are available. The plots of exemplary raw data in figure O.20 and deduced Z and Y data in figure O.21 show how characterisation data can be made comparable even when gained from different driving setups (with and without transformer). This also enlarges the amount of data sets suitable for a comparison of admittance circles, and even the restriction to a single hydrophone position leaves several recordings in the set. Interestingly, admittance circles of many different sizes, they are depicted in figure O.22, can be generated even from this data subset, and the spread of the entire data set (any hydrophone position) is not larger than that. In order to represent the whole span of working conditions encountered with resonator 5, the results of analysing the smallest (case 140), the largest (case 149), and one intermediate (case 187) admittance circle are given in table O.4. The second resonance visible in figures O.20 and O.21 a little below 20 kHz seems relatively minor in the electrical data, but it is in fact the resonance that can be identified with the SF experiment working point described by Saglime [390]. Similar analysis results for that resonance are listed in table O.5. The fitted admittance and impedance circles on the data background can all be found in appendix P.6 in figures P.13, P.14, and P.15.

O.4.2 Acoustic properties

Lastly, the acoustic data recorded on resonator 5 will be presented: sound pressure data recorded with the hydrophone and displacement data recorded with the pickup needle. The sound pressure gain map, spanning the whole distance from the front surface of the bottom piston up to the free surface of the liquid with a spatial resolution of 5 mm is given in figure O.23. The map reveals three resonances in the scanned frequency interval. One cannot simply identify them with the fundamental mode and two subsequent harmonics because even the mode shape of the first resonance has a node at 7.5 cm. Secondly, the positions of the pressure nodes on the vertical axis divide each mode into pieces of unequal length. Thirdly, the lower boundary of the cylindrical main liquid volume is not truly a sound pressure node unlike the free surface at the top, which can be attributed to the acoustic properties of the lower piston. This all makes it clear that the vibration mode shape of the

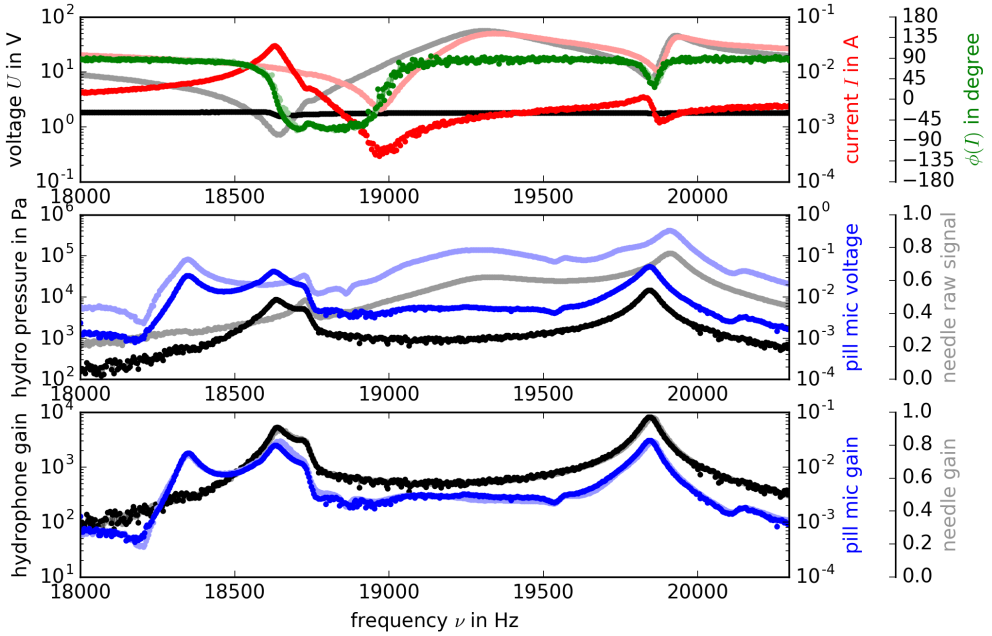


Figure O.20 Raw data of resonator Nº 5: with and without transformer.
The recordings of voltage, current and the phase in between them in the top diagram shows the large impact of the transformer (the case with transformer (case 129) is shaded, the one without (case 131) is shown fully coloured in the foreground). In the middle the transformer influence can be seen by the change of the hydrophone and microphone signals. The congruent data sets in the bottom diagram, however, show that different data sets are very well comparable if one looks at the gain signals. Similarly, looking at Y and Z allows comparisons of electrical data independent from the setup, visible in figure O.21, which represents the same underlying two data sets.

Table O.4 BVD circuit properties in comparison: resonators 5 & 8.
This table shows the admittance and impedance circle analysis results (see the circles in figures P.13 and P.14) gained in four exemplary cases from resonator 5 and compares them to resonator 8 (case 25). The different levels of damping can be directly read from $G_{\max}$, R, or Q_m. The hydrophone had been held ever by the specialised aluminium top head except in case 187, where it had been suspended by its cable and the top head had been taken away. Its position in resonator 5 was always 3.5 cm. Of the alternative values of k the one based on f_s and f_p was used for the figure of merit M. It can be seen that the only numbers needed from the impedance circles are f_a and f_p. In most cases the setup without the transformer was used because it kept the voltage almost stable. However, as this setup leads to very low current amplitudes in the case of a high-Q resonator and a much degraded current phase measurement at the antiresonance, Z-circles of low quality are the consequence. Case 129 is the exception here where the transformer had been in use leading to much nicer Z-circles. It has to be kept in mind that an offset of a few Hertz on each of the antiresonance frequencies $f_a \approx f_p \approx f_n$ does not matter so much because the Δf between the resonance and the antiresonance is much larger in the case of a high-Q system than the deltas found within each group. It also has to be noted that the angle function smoothing routine described in appendix P.4 was applied in each of the cases with degraded phase information. The smoothed phase functions are depicted in figure P.16.

quantity	unit	formula	c. 25	c. 129	c. 140	c. 149	c. 187
T	°C		–	–	–	20.8	19.6
f_{mB}	Hz		17921.7	18608.2	18644.2	18577.7	18677.5
f_m	Hz		18054.7	18639.3	18677.9	18587.1	18688.8
f_s	Hz		18105.0	18642.0	18680.6	18587.4	18689.4
f_r	Hz		18161.2	18644.7	18683.3	18587.7	18690.0
f_{nB}	Hz		18225.5	18673.8	18712.4	18597.3	18701.2
f_a	Hz		18342.6	18970.5	19012.8	18929.4	19086.2
f_p	Hz		18381.4	18974.7	19022.4	18932.0	19087.3
$G_{\max}$	mS		2.47	16.17	13.48	54.0	33.4
B_s	mS		0.837	1.25	1.03	1.58	1.62
Q_m		$\frac{f_s}{f_{nB}-f_{mB}}$	59.6	284	274	945	787
Q_e		$\frac{B_s}{G_{\max}}$	0.34	0.077	0.076	0.029	0.049
R	Ω	$\frac{1}{G_{\max}}$	405.4	61.8	74.2	18.5	29.9
L	mH	$\frac{Q_m R}{\omega_s}$	212.3	150.1	173.2	149.7	200.6
C	nF	$\frac{1}{Q_m R \omega_s}$	0.364	0.485	0.419	0.490	0.362
C_0	nF	$\frac{f_r^2}{f_a^2-f_r^2}C$	18.13	13.77	11.78	13.20	8.44
C_0	nF	$\approx \frac{B_s}{\omega_s}$	7.35	10.64	8.76	13.55	13.83
k		$\sqrt{\frac{f_p^2-f_s^2}{f_p^2}}$	0.173	0.186	0.189	0.190	0.203
k		$\sqrt{\frac{f_a^2-f_r^2}{f_a^2}}$	0.140	0.185	0.185	0.189	0.203
M		$\frac{k^2 Q_m}{1-k^2}$	1.83	10.24	10.12	35.4	33.9
Γ	nF m^{-1}	$\frac{Cd}{A}$	0.204	0.273	0.235	0.275	0.203

Table O.5 Resonator № 5: BVD circuit properties at the second resonance.
This table shows the quantities which can be gained from deducing BVD circuit models from the admittance and impedance circles of the second resonance exhibited by the resonator slightly below 20 kHz. Some formulae differ from table O.4 because no frequencies f_r and f_a can be determined from Y- and Z-circles which are not intersecting with the real axis (see figures P.13 and P.15).

quantity	unit	formula	c. 129	c. 140	c. 149	c. 187
T	°C		–	–	20.8	19.6
f_{mB}	Hz		19819.2	19855.7	19786.6	19933.8
f_m	Hz		19832.8	19868.5	19795.0	19944.9
f_s	Hz		19850.5	19881.6	19804.8	19955.8
f_r	Hz		–	–	–	–
f_{nB}	Hz		19873.1	19901.4	19824.9	19989.6
f_a	Hz		–	–	–	–
f_p	Hz		19873.7	19903.6	19835.9	19981.3
$G_{\max}$	mS		1.21	1.32	1.87	1.71
B_s	mS		1.14	1.12	1.29	1.21
Q_m		$\frac{f_s}{f_{nB}-f_{mB}}$	369	435	516	486
Q_e		$\frac{B_s}{G_{\max}}$	0.941	0.854	0.691	0.711
R	Ω	$\frac{1}{G_{\max}}$	824.0	760.4	535.3	586.4
L	mH	$\frac{Q_m R}{\omega_s}$	2435	2647	2219	2274
C	nF	$\frac{1}{Q_m R \omega_s}$	0.0264	0.0242	0.0291	0.0280
C_0	nF	$\frac{f_s^2}{f_p^2-f_s^2}C$	11.30	10.94	9.25	10.94
C_0	nF	$\approx \frac{B_s}{\omega_s}$	9.15	9.00	10.37	9.67
k		$\sqrt{\frac{f_p^2-f_s^2}{f_p^2}}$	0.0483	0.0470	0.0560	0.0505
M		$\frac{k^2 Q_m}{1-k^2}$	0.861	0.962	1.62	1.24
Γ	nF m^{-1}	$\frac{Cd}{A}$	0.0148	0.0136	0.0163	0.0157

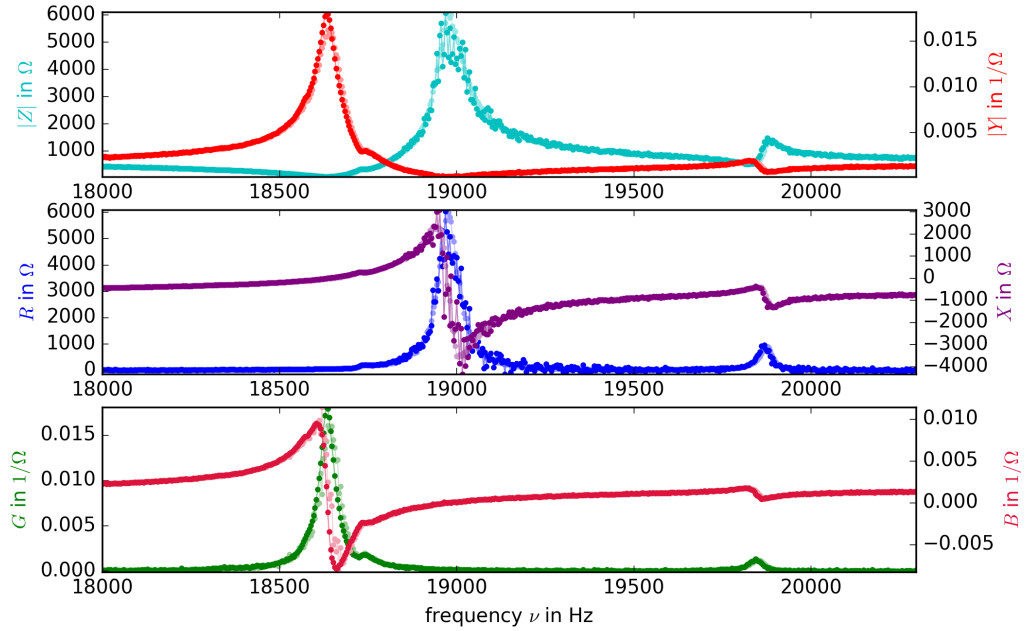


Figure O.21 Z and Y data of resonator № 5: with and without transformer.
This plot is based on the raw data shown in figure O.20 and compares the two cases of driving resonator 5 with (shaded, in background) and without the transformer (full colours, foreground). While the raw signals differ drastically, the pairs of plot lines are congruent here, which illustrates the suitability of the impedance and admittance data for comparisons.

liquid is strongly influenced by the solid structural parts of the resonator. The three sound pressure mode shapes are depicted in figure O.24.

In the latest phase of the measurement campaign the pickup needle for tracking radial displacements was added to the setup (case 206 and later). On that occasion, the Labview data acquisition program was updated from the state depicted in figure P.1 to the state of figure P.2. This made phase recordings available for all analysed signals, as described in appendix P.1.3. The sound pressure mapping had then been repeated, but only covering the range of hydrophone positions possible with the aluminium top head used for hydrophone fixation and therefore lacking the lower part of the mode shapes. These renewed sound pressure amplitude and phase maps are shown in appendix P.6 and they exhibit the interesting feature that the resonance visible in figure O.23 at 19.8 kHz as a weak background resonance has a much stronger interaction with the first liquid resonance in figure P.21.

Analogue glass wall displacement amplitude and phase maps compiled from frequency sweeps are presented in figure O.25 and they reveal how different the displacement mode shapes of the glass wall are from the pressure mode shapes in the liquid. The mode shape profiles are shown in figure O.26. It is interesting that the 1st and the 2rd resonance are multi-belly shapes while the 2nd resonance exhibits an exceptionally wide region ranging on the vertical axis from 0 to 8 cm swinging all in phase.

Looking at the hydrophone and the pickup needle data, considering amplitude

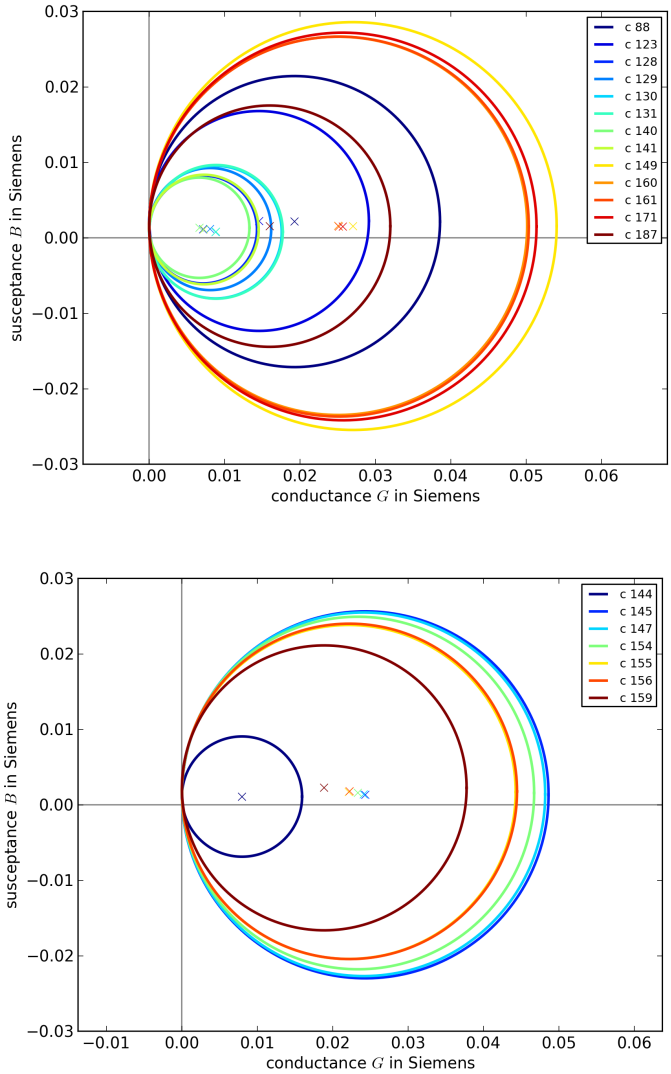


Figure O.22 Admittance circles of resonator Nº 5.
The fitted admittance circles in the diagram correspond to measurements taken over a time span of 16 months. Always the filling level of acetone in the resonator (measured before the introduction of the hydrophone) had been 8.0 cm and the hydrophone was fixed by the aluminium top head at 3.5 cm with the single exception of case 187 where it was hanging at the same position suspended by the cable with the aluminium top head taken away. Astonishingly, it is not that case where the largest $G_{\max} = 1/R$ reveals the lowest level of damping, but case 149 instead. While the whole set covers 16 months, cases 123-187 cover 6 months, 128-141 correspond to three days, and 160-187 to two days. The circles from the sets 145-159 in the bottom diagram correspond to measurements taken in close sequence on one single day. Case 144 had been recorded eleven days earlier. Between cases 144 and 145 the equipment had not been touched except for the refilling of 8 ml of acetone to restore the level and 15 minutes of degassing. Between cases 154 and 155 3 ml of acetone were added to bring the level back up by 0.5 mm to make the flat part of the liquid surface coincide again with the upper edge of the 8 cm pen mark. Between cases 156 an 159 the resonator was cooled down from 21.9 °C to 18.1 °C. This interval is larger than the logged temperature differences experienced in the lab over longer times. The bolts of the top head were not touched during all that time because the refilling could occur with a funnel through the hydrophone outlet. The qualitative changes of the raw signals represented by the data set pairs 144-145, 154-155, and 156-159 can be seen in figures P.18, P.19, and P.20 in appendix P.6. (Between cases 156 and 159 the chamber was left untouched, the data sets recorded in the meantime were for examining the unloaded transducer.) This means that neither the filling level modification nor the temperature change has been able to induce a change in working conditions as substantial as the one having occurred between cases 144 and 145. By consequence, these variables seem unsuitable to explain the varying vibration behaviour of the resonator. The coupling with the aluminium top head cannot be excluded as a source of the variation, but the fact that case 187 is not an outlier on the high-Q side speaks against simply equating the mounting of the top head with the addition of a damping term.

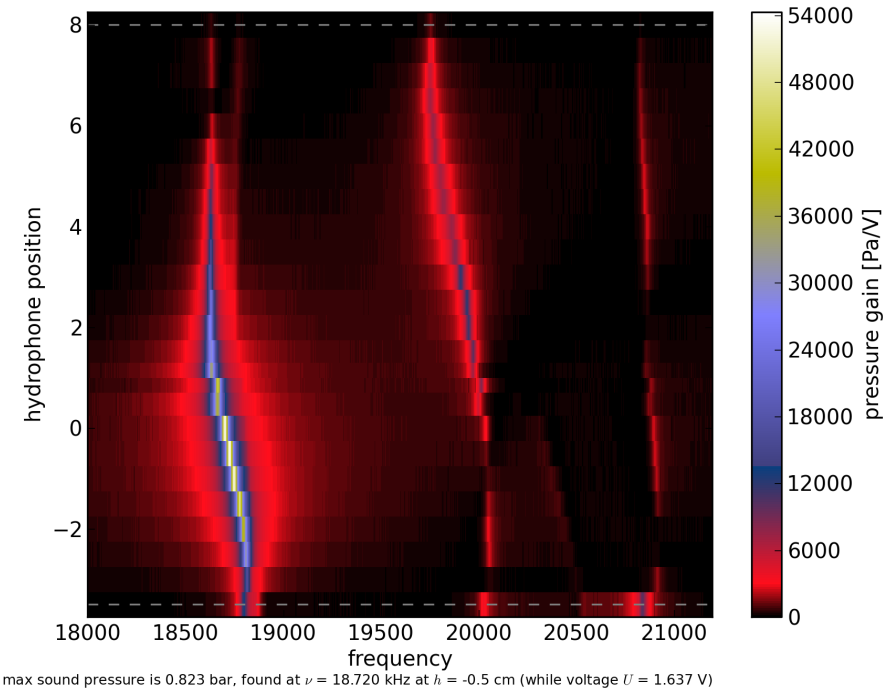


Figure O.23 Sound pressure map of resonator № 5.
This colour map shows the sound pressure gain of the hydrophone in resonator 5. The map incorporates data from 25 frequency sweeps (cases 162-186) recorded with different hydrophone positions. The sound pressure was probed along the central axis from close to the bottom piston's front plate (at a distance of $\approx 0.25\,\mathrm{mm}$) up to the liquid's free surface in steps of 5 mm. The comparison with figure I.1 (representing data recorded by Saglime [390]) shows that the SF trials by the RPI team had been conducted while keeping the resonator in the second depicted resonance. The cavitation position statistics in [496, 497] indicate that also Taleyarkhan's Purdue team targeted this mode shape. The map reveals nicely how the hydrophone being pushed down into the liquid and replacing it tunes the resonance frequencies upwards. It seems that the frequency shift is stronger when the hydrophone tip traverses a sound pressure antinode as compared to when it is near a sound pressure node.

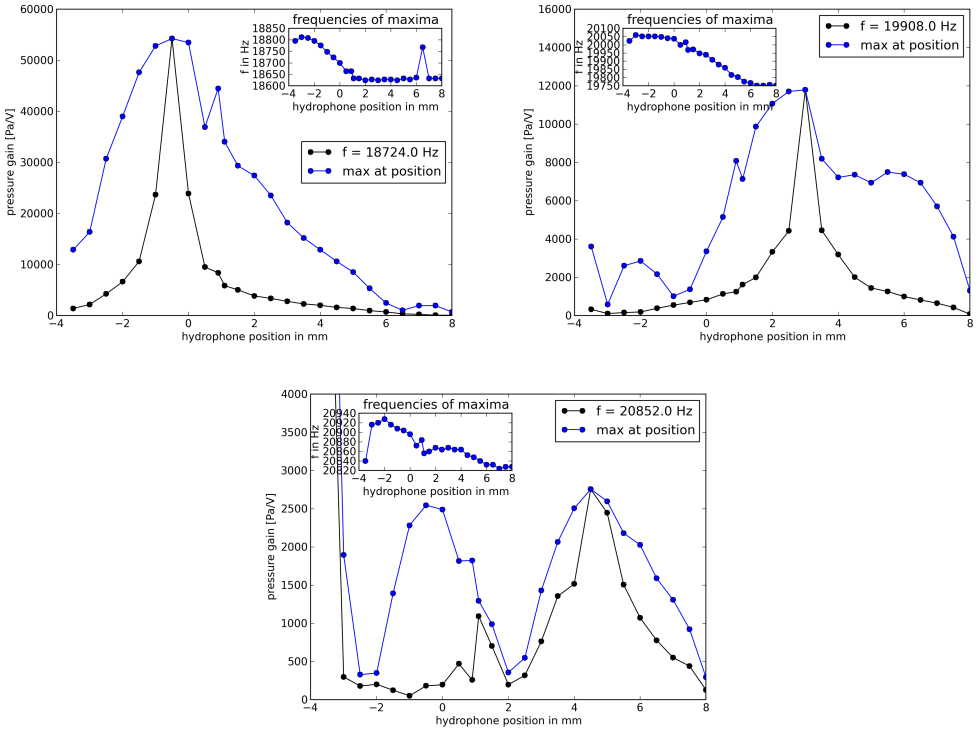


Figure O.24 Sound pressure profiles of resonator Nº 5.
The three profile pairs shown above correspond to the three resonances in the dataset shown in figure O.23 at 19.7, 20, and 21 kHz. As one can see by the changes between the blue and the black profiles, it makes a very big difference in the case of resonator 5 with its high Q-factor whether the hydrophone position is simply varied at a constant frequency or whether the local sound pressure maxima along the frequency axis are concatenated. Only the latter approach under the condition of a sufficient frequency resolution (4 Hz in this case) results in a meaningful mapping of the mode shapes of the standing sound pressure waves. The comparison with the reprinted measurement of Saglime [390] in figure I.1 shows that the SF trials by the RPI team had been conducted while keeping the resonator in the second depicted resonance. The underlying dataset (cases 162-187) has been recorded on two days. The zigzag visible in the blue profile of the lowest resonance (top left) hints towards the transition between the two halves of the set. The first subset (cases 162-177) corresponds to the hydrophone positions $[8.0, 7.5, 7.0, ..., 1.0, 0.5]$ in centimetres and the second day's subset (cases 178-187) to $[1.0, 0.0, -0.5, -1.0, ..., -3.5]$. In the plots of this figure and the pressure gain map of figure O.23 the twice occurring data from 1.0 cm are depicted at 1.1 (case 176) and 0.9 cm (178). In order to be able to reach the lower positions with the hydrophone, the aluminium top head had to be taken away between cases 177 and 178. From then on the hydrophone has been suspended only by its cable. The sound pressure increase from the first to the second set may thus be due to the structure being able to vibrate more freely without the damping through the rubber ring and bolts connection to the top head, but it may also have originated from the instability over time of the resonator state observed throughout the measurement campaign and illustrated in figure O.22. The profile of the second mode might also be affected from unstable working condition and should be compared to the counterpart in figure P.22. Considering the lowered values between 3.5 and 6 cm forming a marked dent in the mode shape data, cavitation in front of the hydrophone tip might be an explanation. The pressure gain of that resonance of about $\approx 0.1\ \frac{\text{bar}}{\text{V}}$ compares to $\approx 0.05\ \frac{\text{bar}}{\text{V}}$ reported by Saglime [390].

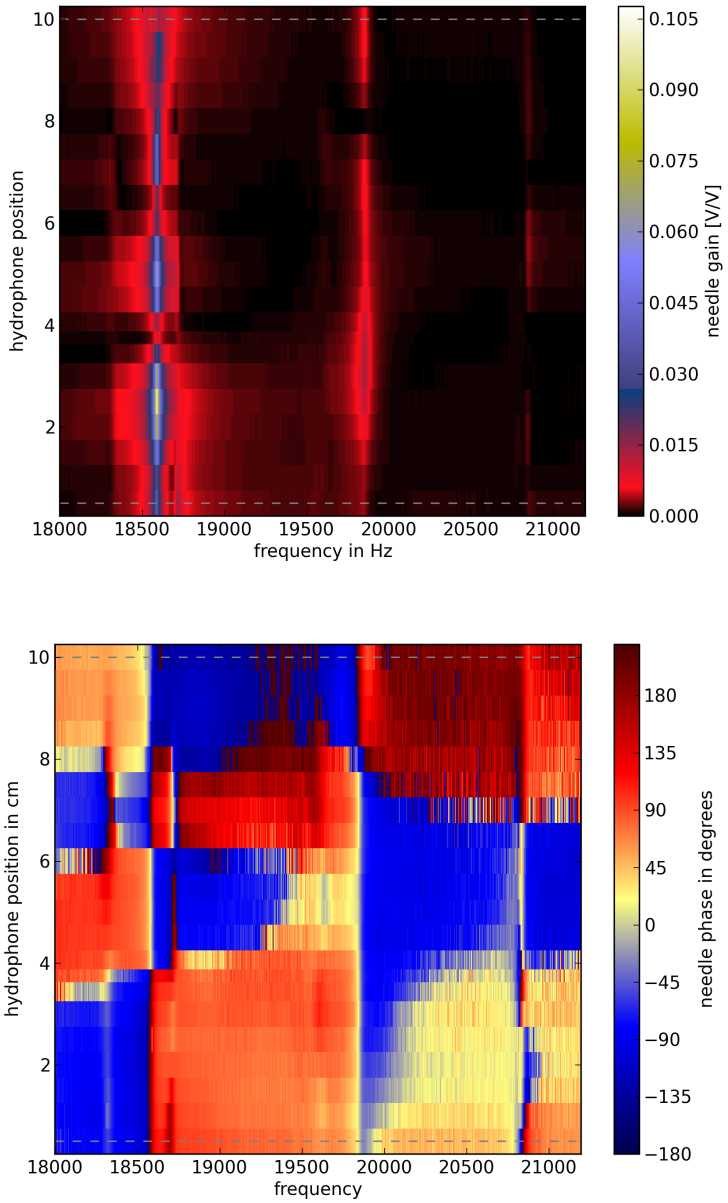


Figure O.25 Pickup needle displacement map of resonator Nº 5.
The colour maps above show the amplitude of the pickup needle gain and the phase of the pickup needle signal. The dataset represents cases 257-282. Only the glass wall above the piezoelectric transducer was scanned. The highest pickup needle position (10 cm) is shortly below the aluminium flange with its slightly protruding silicone bead, and the lowest position is 5 mm from the transducer top edge. The acetone filling level is at 8 cm, for the 2nd resonance this position corresponds to a displacement node visible in the amplitude map. The phase map shows that for both the 1st and the 2nd resonance the 8 cm-line coincides with blue-red borders. On the 3rd resonance no shape difference can be made out between above and below the liquid surface.

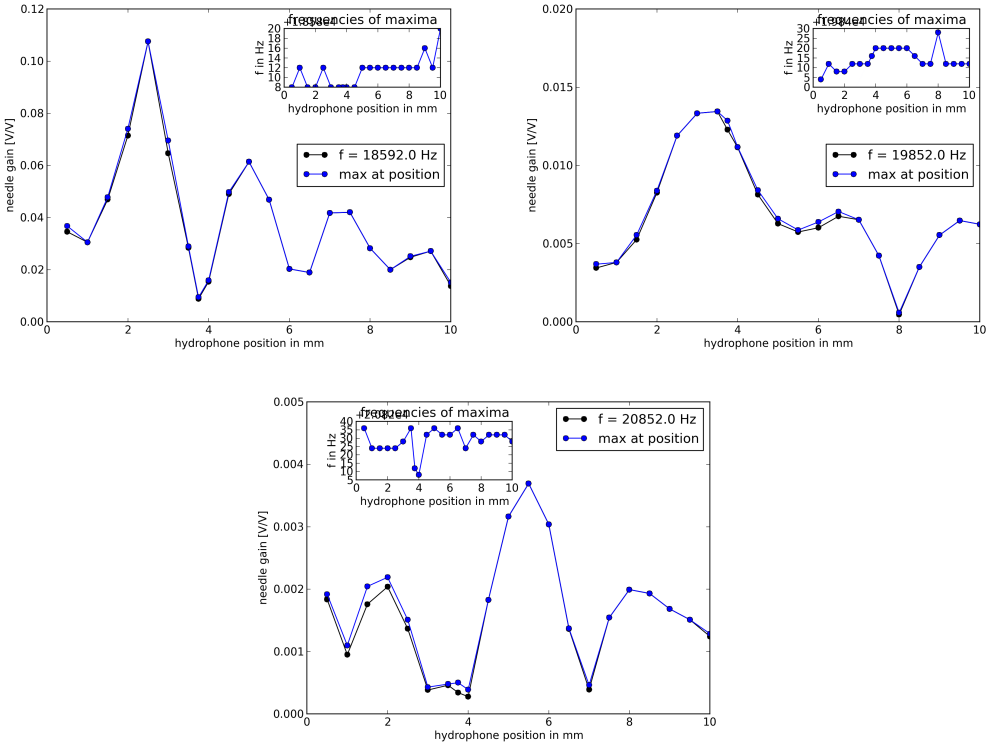


Figure O.26 Pickup needle displacement profiles of resonator Nº 5.
The above plots show the vertical pickup needle gain profiles of the three resonances exhibited by the resonator in the scanned interval. As expected in this case where the position of the measurement device does not influence the resonator, the difference between the vertical cuts (black) and the concatenation of local maxima (blue) is negligible and the frequency ranges covered by the insets are narrow.

and phase in each case, from the point of view of the FEM simulations needing validation data, three important remarks can be made: (a) the displacement data is richer in detailed structure, and (b1) an equivalent of the displacement map can be gained with one single FEM simulation, whereas many simulations with a varying hydrophone tip position are needed to create the equivalent of the sound pressure map, or (b2) reading a pressure map from one single FEM simulation where the hydrophone has not been modelled means having to deal with an additional contribution to any map mismatch. However, (c) the hydrophone data comes with a calibration function translating from Volts into Pascal, whereas for the entire displacement sensor, involving the unspecified magnet-coil transducer in combination with the amplifier and the resistors at its output, no calibration experiment could be made available so far.

The data presentation is rounded off by figure O.27 showing Q-factors which can be deduced from the pressure and displacement sweeps and figure O.28 compiling pressure gain observations for the 1^{st} and 2^{nd} resonance.

O.5 Summary

O.5.1 Shortcomings and improvement possibilities

It lies in the nature of finite measurement campaigns that the generated data collection is not as systematic and complete as may be desired, that the type of measurements and the equipment setup reflects the status of knowledge and questions having existed at the time. It makes sense to list weaknesses and improvement options to improve the starting position of follow-up projects. Next to obvious and profane measures like a finer frequency resolution around resonances or better adjustment of the current probe resolution some important points would be

- collecting data on the dependency of Q-factors on the driving power,
- adding an independent calibration experiment for the displacement pickup needle,
- making full use of the 2D tracking capability of the pickup needle on roughened surfaces and collecting 3D motion data by combining different probe orientations,
- expanding the displacement measurement to larger surface areas of the resonators and to the outside electrode of the transducer (which can be the one on ground potential),
- examining the vibration behaviour of single assembly parts (like pistons) separately,
- examining other simple plate, beam, or cylinder parts for more validation data and the purpose of material constant determination (glass, PZT, epoxy, RTV silicone),

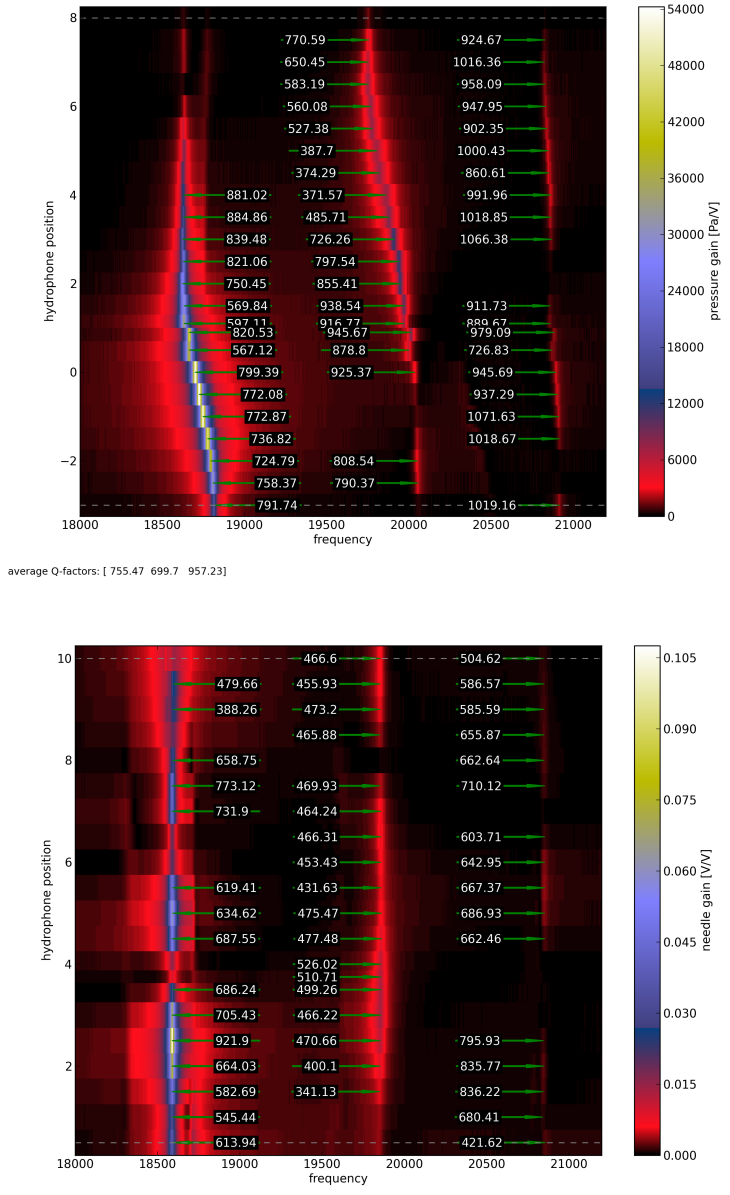


Figure O.27 Resonator № 5: acoustic Q-factors.
These are the same amplitude maps of the hydrophone (top) and the pickup needle signal (bottom) as in figures O.23 and O.25, but this time with annotations containing Q-factors computed by the formula in table J.2. For each of the three resonances the maximum amplitude was scanned first, and the peaks at other vertical positions were only analysed if they reached at least 20 % of that maximum height. Other exclusion criteria were side peaks above $1/\sqrt{2}$ times the peak height and substantial asymmetry. The average Q-factors for the three resonances are indicated in each plot's lower left corner.

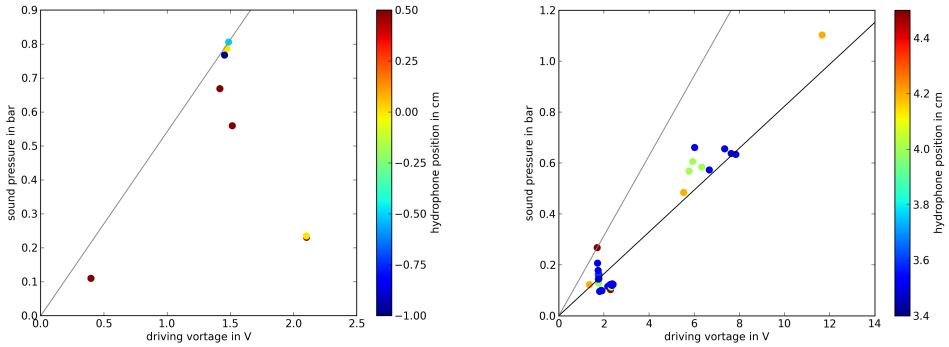


Figure O.28 Resonator Nº 5: maximum pressure response.
These plots aggregate hydrophone-measured peak sound pressure per driving voltage information for the first resonance at 18.6 kHz (left) and the second one slightly below 20 kHz (right). In the case of the first resonance the pressure antinode is found around a hydrophone position of 0 cm where not many frequency sweeps have been recorded. The point cloud represents peak values found between 18.5 and 19 kHz for all cases where the hydrophone position was in the range from −1.0 to 0.5 cm. The grey line marks the maximum pressure gain of 0.543 $\frac{\text{bar}}{\text{V}}$ found in case 180. In the case of the second resonance hydrophone positions from 3.4 to 4.5 cm were taken into account and the max gain (grey line) was 0.157 $\frac{\text{bar}}{\text{V}}$ (case 150). For this larger data set of 58 sweeps a linear fit forced through the origin is also given by the black line. It corresponds to a gain of 0.0823 $\frac{\text{bar}}{\text{V}}$ (error: 2.5×10^{-3} $\frac{\text{bar}}{\text{V}}$, R-value: 0.969). All these values are about an order of magnitude larger than the response of resonator 8.

- using smaller sound pressure probes with less impact on the sound field, e. g. a diaphragm-based extrinsic Fabry-Perot interferometric optical fiber sensor [189], which, because of its mounting on a glass fibre, can be bent around obstacles like an installed upper piston,
- using strain gauges,
- or scanning surface displacements with laser interferometry.

O.5.2 Conclusions

In this appendix chapter the postprocessed data of a measurement campaign for characterising SF resonators carried out at the Gaerttner Laboratory at RPI was presented. Although several aspects in which the data is still suboptimal may be pointed out, it nevertheless represents a substantial enhancement of the knowledge about SF resonators collected in the preceding works. It is helpful in better understanding the oscillation dynamics and suitable for benchmarking resonator simulations.

The electric data, admittance and impedance circles, give direct access to an integral measure of damping and allow to include a bare transducer ring into the comparison. The pre-existing setup for recording acoustic data, formerly restricted to microphone and hydrophone signals, was extended to include structure displacement data. That data is collected with a simple pickup needle in a way not interfering with the vibration mode shape. The displacement data exhibits a richness in structure and information content being of valuable advantage for understanding the resonators' vibration behaviour and validating simulations of it.

The design of resonator № 8, the adaption of the resonator layout by Cancelos to SF experiments, manifests the attempt of the RPI team to come up with an ab initio design solution for the SF resonator design problem in a way enabling the repeated manufacturing of the same resonator type with small geometric tolerances and a well-defined performance. The experimental campaign described here took up the work and began with the characterisation of resonator 8 and first cavitation trials with it. Although it could be verified that the resonator is able to induce the intended sound pressure mode shape in the liquid, these experiments revealed a substantial drawback of the design in the form of too much damping. The dissipation of vibration energy turned out to be so high, that it is much less suitable for repeating the SF experiment than the previous resonator versions. The reasons may be the choice of materials, a disadvantageous geometry, or the way of connecting assembly parts. The aluminium flanges are assumed to play an important role. Another drawback of the bolted flanges is that they make reliable FEM simulations exceedingly difficult. The incremental simulation-aided optimisation of such a resonator design is therefore no valid option.

The focus was then turned back to the West-Howlett design of SF resonators, but not for a repetition of the failed SF trials with the old equipment and its persisting issues (e. g. acoustic performance sensitivity and instability over time), but for gathering validation data, the 2D sound pressure and displacement maps in figures O.23, O.25, and P.21 most prominently, and for serving the goal of coming up with a next generation of SF resonator designs in a simulation-aided ab initio design process. The Q-factor and the sound pressure per driving voltage of resonator № 5 are taken as the highscore to be at least matched in simulations by any new design proposal.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
A	surface area
B	susceptance
C	capacity
$\mathbb{C}$	complex numbers
d	transducer thickness
$\vec{E}$	electric field
f	frequency
G	conductance
h	transducer height
I	current
k	electromechanical coupling coefficient
L	inductance
M	figure of merit
Q	quality (“pointedness” of a resonance peak)
Q_e, Q_m	electric and mechanical Q-factor

R	resistance
$\mathbb{R}$	real numbers
r	radius; capacitance ratio
U	voltage
X	reactance
Y	admittance
Z	impedance

List of Greek quantity symbols

Symbol	Description
Γ	motional capacitance constant
ν	frequency
ϕ, φ	phase angle
ω	angular frequency

List of abbreviations

Abbreviation	Description
ADC	analogue-digitial conversion/converter
alu	aluminium
amp	amplifier
BNC	Bayonet Neill–Concelman (coaxial cable connector)
DAC	digital-analogue conversion/converter
FE,FEM	finite element (method)
FFT	fast Fourier transformation
GPIB	General Purpose Interface Bus (IEEE bus specification standard)
HiFi	high fidelity (high-quality audiophile stereophonic sound reproduction)
HP	Hewlett-Packard®
hydro	hydrophone
IEEE	Institute of Electrical and Electronics Engineers
mic	microphone
MM	moving magnet
NI	National Instruments®
OD	outer diameter
PC	personal computer
PCB	short for picocoulomb (in the company name PCB Piezotronics®)
PCI	Peripheral Component Interconnect (local computer bus standard)
PuBe	Plutonium-Beryllium (mixture used as noutron source)
PZT	lead zirconate titanate (a piezoelectric ceramic)
RPI	Rensselaer Polytechnic Institute
RTV	room temperature-vulcanising (silicone)
SF	sonofusion

Appendix P

Additional documentation on resonator characterisation

P.1 Labview codes

Labview® is a software for integrating laboratory equipment with a computer. It offers an environment for visually programming applications for either just monitoring and recording incoming signals, or also controlling laboratory equipment via outgoing signals. Executable applications programmed in Labview are called "virtual instruments" or short "VIs". VIs are composed mainly of loops, basic logic and other VIs, then called sub-VIs. Within a complex VI with many sub-VIs one finds in general a mixture of on the one hand VIs which are part of original software packages and on the other hand VIs created by users.

P.1.1 Pre-existing applications

The list of relevant Labview VIs (`*.vi`) created by researchers at RPI includes:

`BF Control V2.vi` contains a control loop for adjusting the excitation frequency to stay in resonance (functions by aiming to keep the phase of either the mic or the transducer signal close to a target value chosen through the GUI; the control is of PI type); determines bubble burst rates from mic and transducer signals and plots their time-development

`BF Temperature Control V2.vi` same as above with basic control loop for fan and freezer activity

P.1.2 Applications developed for this project

The list of VIs created in the course of the cooperative project of RPI and KIT on bubble fusion includes VIs authored by Bernie Malouin and Markus Stokmaier (with prefix "BFBM") and ones authored by Markus Stokmaier (prefix "BFM"). All these VIs sweep through a range of driving frequencies step by step, stabilise a stationary oscillation of the resonator and record electrical and acoustic properties of the piezo-driven SF resonator.

`BFBM_chamber_characterisation.vi` and its derivatives were used for recording data sets 1 through 10. The amplitude and phase of signals are determined through these steps (carried over from `BF Control V2.vi`): (a) multiplication of a signal snippet with a Hanning window, (b) transformation of the signal $s(t) \rightarrow \hat{s}(f)$ by fast Fourier transform (FFT), (c) detection of point f_{res} of maximum $|\hat{s}(f)|$, (d) amplitude and phase are real and imaginary parts of $\hat{s}(f_{\mathrm{res}})$.

`BFBM_chamber_characterisation_edit_d*.vi` and later (case 11 and later): frequency and amplitude of dominating harmonic signal determined by the sub-VI `Extract Single Tone Information.vi` of the library `NI_MAPro`.

`BFM_chamber_charact_v_1_90.vi` (case 190 and later): this version and older ones use scopes A & B for ADC, they rely on the sub-VI `AI Acquire Waveforms.vi` (from the library `AI.llb` for analogue input) to communicate with the scopes via GPIB; the scopes must be externally triggered by the function generator

`BFM_chamber_charact_v_2_00.vi` (case 206 and later): from this version on signal ADC was accomplished with the NI PCI-6025E card mounted inside the control computer employing the scheme shown in figure P.2; in this setup the ADC card needs to have the function generator signal available for triggering (in the diagram it can be seen that input `PFI 0` of the ADC unit was used for triggering)

The laboratory log book, where the whole measurement campaign has been documented, gives information on which VI has been used for which measurements.

P.1.3 Key details inside used Labview VIs

Detail A1 – waveform analysis, early version: The goal is to determine amplitude and phase of digitised sinusoidal voltage signals. In this version the analogue-to-digital conversion is made by two GPIB-capable oscilloscopes. Their signal buffers are sent to the PC where they are analysed. The time and amplitude axis resolutions of the scopes have to be kept within suitable ranges by using the manual switches in order to ensure a decent quality of the buffered data. Figure P.1 shows the Labview structures used for the data reading and analysis. The sub-VI labelled "AMP. FREQ." with the full name `Extract Single Tone Information.vi` extracts the amplitude and phase information. Only one phase signal, the phase of the current relative to the voltage, is being stored. Voltage, hydrophone, and microphone data are stored as amplitudes without phase information.

Detail A2 – waveform analysis, later version: The newer setup is shown in figure P.2. Its core is made up of the four sub-VIs for the preparation and the execution of the A-D conversion of an analogue input (AI) using the NI® PCI-6025E card. At the end of this row follows the amplitude and phase extraction with the same sub-VI "AMP. FREQ.". In this setup (used from case 200 and onwards) there is also a slot for the pickup needle data. The second improvement is that phase information is not only stored for the current, but for all other signals as well. Unfortunately, because of a wrong trigger setting during the short final part of the

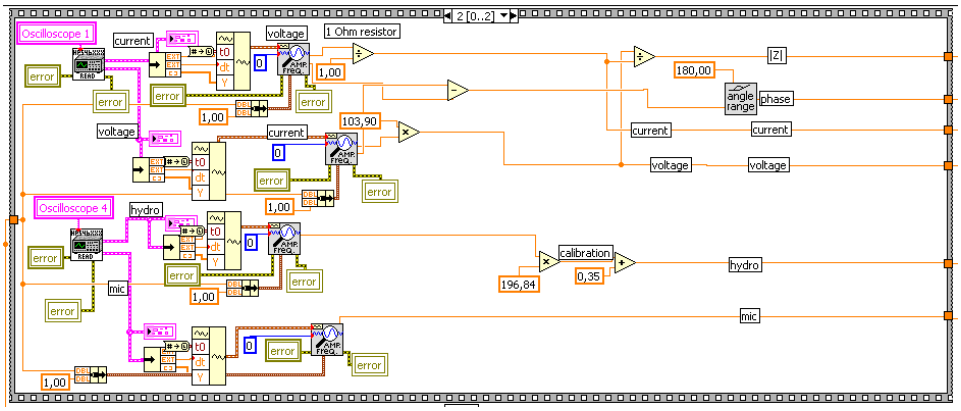


Figure P.1 The sub-VIs used for waveform analysis (early version):
In this setup, two GPIB-capable oscilloscopes serve as A-D converters. They are represented symbolically on the left. Two signal lines can be extracted from each scope. The signal analysis is accomplished by the sub-VI labelled "AMP. FREQ." with the full name `NI_MAPro.lvlib:Extract Single Tone Information.vi` indicating that it comes from the library `NI_MAPro`. This subroutine executes multiple signal treatment operations like windowing, FFT, frequency range selection, peak detection. Its output values are amplitude, frequency, and phase in degrees. In the case of a known frequency or frequency range, the search space of the sub-VI can be narrowed in order to get a more accurate or reliable amplitude and phase reading from a not perfectly clean signal. This can be done by specifying the interval of interest through its centre frequency and width (in percent of the sample rate) and feeding that value pair into the sub-VI from below. (The value passed as centre frequency is the same as the one given to the function generator.) In this setup the ADC time and amplitude resolutions depend on the scope settings (in particular the amplitude ranges have to be constantly readjusted manually while sweeping through resonances, some switch events can e. g. be seen as small jumps in the hydrophone data in figure P.8). The conversion visible in the diagram which is applied to the hydrophone signal is based on a mistake, it needs to be reversed as part of the data postprocessing in order to get to useful sound pressure data through the calibration discussed in appendix P.2.

measurement campaign, there is more noise as needed on the phase signals recorded with this setup and they are also shifted. An offset of 40° has been added as a rough correction whenever such data is discussed here. This correction makes the current phase data consistent with the earlier measurements upon a qualitative look and it also rotates admittance and impedance circles into a position where they could be expected, but as the nature of the wrong trigger setting cannot be inferred from the lab documentation, it is impossible to quantitatively analyse the admittance and impedance data from these measurements.

Detail B – compensation factor 2.00 for current data: According to the manual of the current probe amplifier Tektronix® AM 503, the "current/div" switch is calibrated for signals being tracked with an oscilloscope with a setting of 10 mV/div (i.e. millivolt per division). At this setting 1 mV displayed by the scope can be interpreted directly as 1 mA measured by the current probe. The range of currents seen during SF resonator characterisation necessitated a setting of 20 mV/div on the Tektronix® AM 503. The multiplication factor of 2.00 in the Labview code compensates for this. Unfortunately, the Tektronix® AM 503 is not equipped with any GPIB interface, hence the mV/div setting could not be controlled dynamically from within Labview, which might have been a missed chance to improve the signal quality at low currents (in particular at the antiresonance $f_2 \to (f_a, f_p, f_n)$).

Detail C – compensation factor 103.9 for voltage data: As shown in

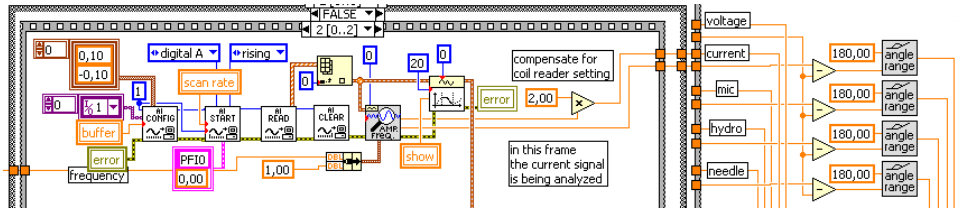


Figure P.2 The sub-VIs used for waveform analysis (later version):
The important sub-VIs are the ones forming the horizontal sequence in the left part of the diagram: `AI CONFIG.vi`, `AI START.vi`, `AI READ.vi`, `AI CLEAR.vi` (all from the library `AI.llb` for analogue input), and `NI_MAPro.lvlib:Extract Single Tone Information.vi` (which carries the label "AMP. FREQ." in the GUI). The first four are necessary for reading out and creating a data sequence from a designated channel of the ADC, the PCI-6025E board in this case. The innermost one of the grey frames is a sequence structure. One after the other, the voltage, current, microphone, hydrophone, and displacement pickup needle signal are analysed and amplitudes and phases passed on through the orange nodes and cables to the right. Treating one signal at a time allows A-D conversion at the maximally available sample rate. (The effective sample rate was 1.05×10^6 samples per second.) In the cases of voltage and current the triangular multiplication node contains the factors 103.9 and 2.00, respectively, for the other signals it is neutral. The regular grid of subtraction nodes and angle range readjustment sub-VIs seen on the right outside the grey frames determines that all signal phases are recorded as values relative to the voltage phase. Therefore, the phase of the voltage signal itself is defined to be zero and not stored. Unfortunately, because of a wrong trigger setting there is more noise as needed on the phase signals recorded with this setup and they are also shifted. An offset of 40° has been added as a rough correction whenever such data is discussed here.

figure O.1, instead of the whole voltage fed to the piezo transducer only a reduced voltage across one of two resistors in series has been read out. The correction factor to reconstruct the original voltage can be calculated based on the two equations

$$\frac{U_1}{U_2} = \frac{R_1}{R_2} \tag{P.1}$$

(where U_i is the voltage drop across resistor R_i) and

$$U_{\mathrm{PZT}} = U_1 + U_2. \tag{P.2}$$

It follows, that

$$U_{\mathrm{PZT}} = U_2 \left(\frac{R_1 + R_2}{R_2} \right) = U_{\mathrm{scope}} \left(\frac{R_1 + R_2}{R_2} \right), \tag{P.3}$$

meaning that the voltage supplied to the transducer is the measured voltage times a compensation factor $\alpha = \frac{R_1+R_2}{R_2}$. The two resistors used in the setup were labelled $R_1 = 100\,\mathrm{k\Omega}$ and $R_2 = 1\,\mathrm{k\Omega}$ with a precision rating of $\pm 2\%$. In reality, their resistances were determined by multimeter to $R_1 = 99.45\,\mathrm{k\Omega} \pm 0.05\,\mathrm{k\Omega}$ and $R_2 = 999.1\,\Omega \pm 0.1\,\Omega$, which would yield $\alpha = 100.54 \pm 0.05$. However, in the running system with the Wheelock amplifier driving the SF resonator at 18-21 kHz with low enough voltage amplitudes and the transformer shorted, the proportionality factor between U_{PZT} and U_2 was measured directly and determined to be 103.9 ± 0.15.

Detail D – wait time after frequency setting: The innermost frame in figure P.2 is a sequence structure. The last of three frames is visible. In the first frame the new desired excitation frequency is communicated via GPIB to the function generator. In the second frame a wait time of 100 milliseconds is specified. Only after that time delay the data acquisition and analysis is executed in the third frame.

The purpose of the wait time is to ensure a steady vibration state of the resonator, i. e. that all transient contributions to the motion pattern caused by the frequency shift (or by switch-on events) have decayed to negligibility. The determination of the delay time to 100 ms was a conservative decision after having inspected the transients recorded after switch-on and switch-off events.

Detail E – conversion of hydrophone voltage to sound pressure: Up to case 191 the hydrophone output voltage has been put through the conversion ($f(x) = ax+b$ with $a=196.84$ and $b=0.35$) which can be seen implemented in figure P.1. That conversion represented the intention to gain the sound pressure amplitude in PSI from the hydrophone signal conditioner output in Volt, but it was based on an erroneous interpretation of the calibration data table. Corresponding data must be converted back via $f^{-1}(x)$ before analysis. From case 192 on the unconverted amplitude has been stored. In appendix P.2 the hydrophone calibration data table and its correct interpretation are discussed.

P.2 Hydrophone calibration

The calibration data of the employed hydrophone as supplied by the manufacturer is listed in table P.1.

Table P.1 Calibration data for the PCB Piezotronics® model S113-A26 hydrophone.

sound pressure [psi]	output voltage [mV]
10	98
20	200
30	301
40	402
50	505

A linear fit of the form $f(x) = Bx$ can be made for that data yielding $B = 99.3833\,\frac{\text{psi}}{\text{V}}$ (shown in figure P.3 on the left), and it can be used to generate the conversion formula for computing the sound pressure in SI units from the hydrophone signal:

$$f(x) = cBx,$$

whereby $c = 6894.757\,28\,\frac{\text{Pa}}{\text{psi}}$ is the conversion from psi to Pascal.

P.3 Details of the electrical analysis of the unloaded transducer

P.3.1 Determining the antiresonance frequency of the unloaded transducer

As can be seen in figure O.8, the phase signal is of low quality in the region around the antiresonance, where the current amplitude and its signal-to-noise ratio are low.

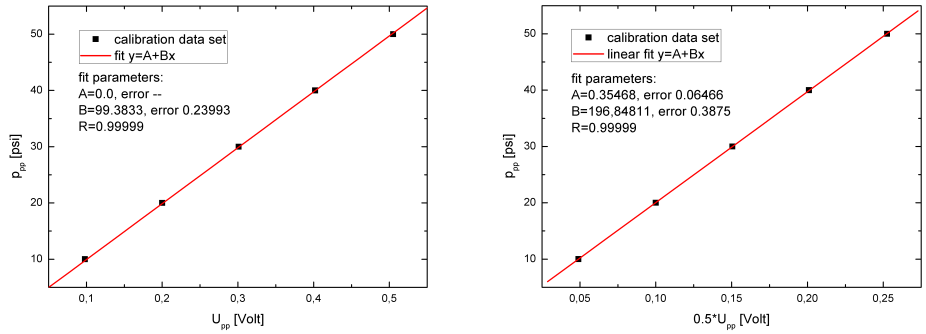


Figure P.3 Hydrophone calibration data.
The left hand plot shows a linear fit forced through the origin of the hydrophone calibration data. That curve can be taken to translate output voltages (in Volt) into sound pressures (in psi). The plot on the right hand side reconstructs the mistake having led to the wrong conversion formula that was applied to characterisation data through case 191.

In order to give a useful estimation of the antiresonance frequency, i. e. the point on the frequency axis where the phase passes through zero, a curve of the form

$$\varphi(f) = \frac{180}{\pi} \arctan\left(\frac{f-\mu}{\sigma}\right)$$

has been fitted to the phase data. The fit parameters μ and σ determine the location and width of the transition region. The best fit, with $\mu = 17\,121.7\,\mathrm{Hz}$ and $\sigma = 31.7\,\mathrm{Hz}$, was found by minimising the sum of square distances of the data points from the model in the interval $[16.5\,\mathrm{kHz}, 18.0\,\mathrm{kHz}]$. The minimisation method was Powell's method as implemented in SciPy [223, 361]. The result is the red curve in figure P.4.

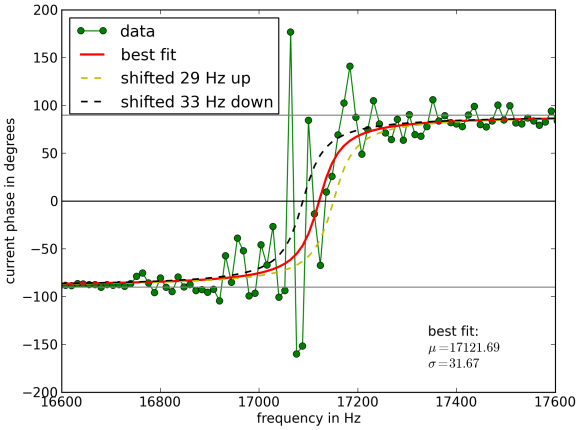


Figure P.4 Case 157 phase data fit.

The antiresonance f_a of the transducer is the frequency where the imaginary part of the impedance $Y = 1/Z$ is zero, and this is equivalent to voltage and current being in phase. Hence, f_a is equivalent to the anchor point $\mu = 17\,121.7\,\mathrm{Hz}$ of the fit function. The dashed lines in figure P.4 show the fit curve after shifts by $-33\,\mathrm{Hz}$

to the left and 29 Hz to the right. The accompanying figure P.5 shows that this corresponds to the limits within which the square distance sum varies only about 10 %. With that backing it is said in appendix chapter O.2.2 that the antiresonance frequency f_a of the examined unloaded transducer is 17 120 Hz ±30 Hz.

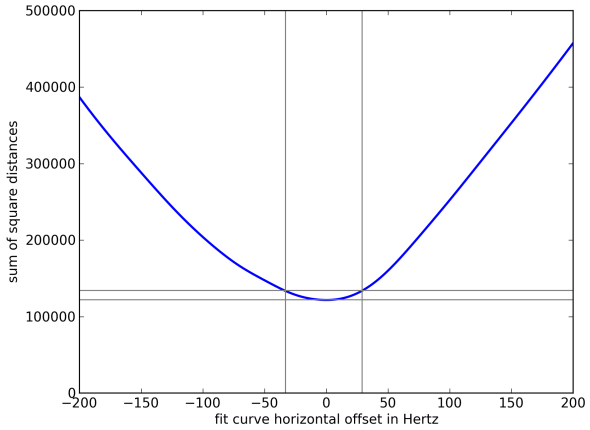


Figure P.5 Case 157 phase fit quality over offset.
The best fit parameters were $\mu = 17\,121.7$ Hz and $\sigma = 31.7$ Hz. This plot shows the objective function of the minimisation over a shift $\Delta\mu$ added to μ, i. e. how the sum of square distances of the data points from the fit changes under variation of the anchor point μ while σ is being kept constant. The upper horizontal grey marker line indicates where the objective function has grown 10 % above the minimal value.

P.3.2 Frequency response of the equivalent circuit

The voltage drop over a chain of several resistors in series can be calculated by summing up all the single resistances. In general, for series structures one has to add up the resistances and for parallel channels the sum has to combine the conductances. For circuits under oscillating voltage loads, it is the impedances that need to be added in the series case and the admittances for a structure of parallel channels. Thus, the total impedance and admittance of simple circuits can be computed with the help of the equations

$$Z_R = R, \qquad Z_L = i\omega L, \qquad Z_C = \frac{1}{i\omega C}, \tag{P.4}$$

and

$$Y_R = \frac{1}{R}, \qquad Y_L = \frac{1}{i\omega L}, \qquad Y_C = i\omega C, \tag{P.5}$$

giving the impedances and admittances for the three component types resistor, inductor, and capacitor. For the simplified equivalent circuit of a piezoelectric transducer given in figure J.2 where the main branch has the total impedance $Z_{\text{main}} = R + i\omega L + 1/i\omega C$ and the branch with the parallel capacitance C_0 has the impedance $Z_{\text{parallel}} = 1/i\omega C_0$, the admittance of the whole circuit is

$$Y_{\text{tot}} = Y_{\text{main}} + Y_{\text{parallel}} = \frac{1}{Z_{\text{main}}} + \frac{1}{Z_{\text{parallel}}} = i\omega C_0 + \frac{1}{R + i\omega L + 1/i\omega C}. \tag{P.6}$$

Plotting $Y_{\text{tot}}(\omega)$ and its inverse, the effective impedance $Z_{\text{tot}}(\omega) = 1/Y_{\text{tot}}(\omega)$ over the Y and Z curves inferred from the measured raw data gives hints about how well the equivalent circuit model can represent the behaviour of the transducer. This is done in figure P.6 including several cases corresponding to various values of C_0 reflecting the different ways to obtain C_0 based on the formula collection in table J.2. The plot shows that combining information from both the resonance and the antiresonance results in an equivalent circuit model which adequately describes the characterised transducer.

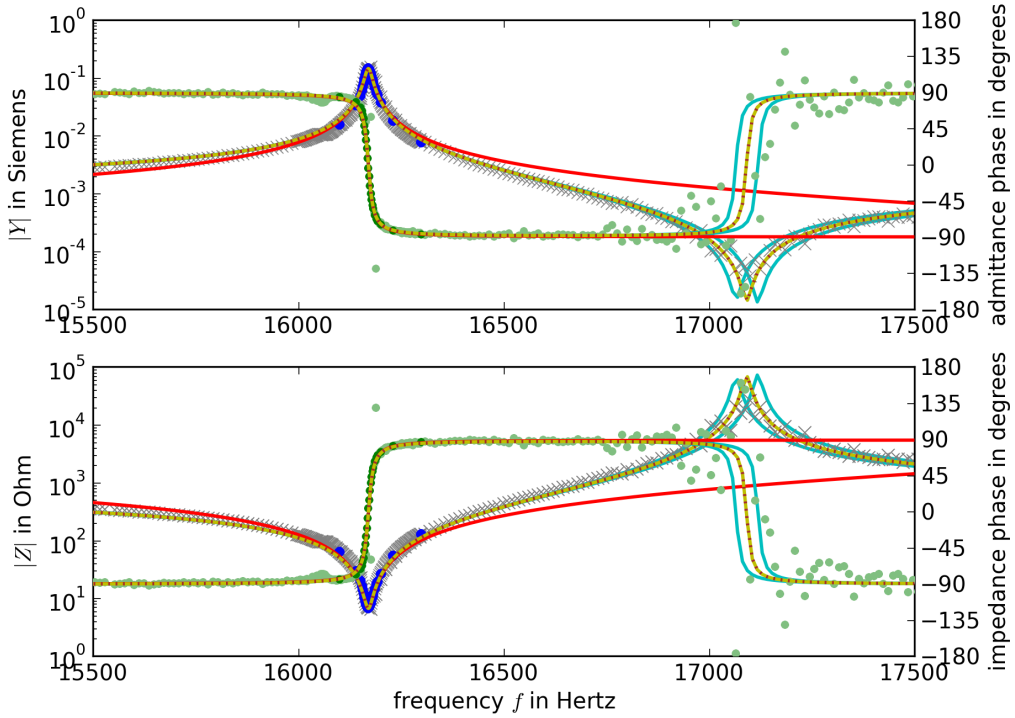


Figure P.6 Free transducer: equivalent circuit model over measured data
These plots show the amplitude and phase of admittance Y (upper plot) and impedance Z (lower plot) of the piezoelectric transducer separated into amplitude (grey crosses, left scales) and phase (green dots, right vertical axes). A subset (fully coloured blue and green dots) of the measured data (cases 157 & 158) has been used for the Y-circle analysis as discussed in appendix O.2. The lines in the plots correspond to several equivalent circuit models (same colours for amplitudes and phases). The equivalent circuit quantities R, L, C can be easily gained by only examining the Y-circle, i. e. the resonance. The last quantity, C_0, can be deduced also from the Y-circle via $C_0 \approx B_s/\omega_s$ or alternatively with information about the antiresonance (f_a, f_p, f_n). This comparison is also given in the plots. The red lines represent the case where $C_0 \approx B_s/\omega_s = 2.9\,\text{nF}$, the yellow line $C_0 = rC = 13.5\,\text{nF}$ with $r = f_r^2/(f_a^2 - f_r^2)$, and the dotted brown line represents the case $C_0 = rC = 13.5\,\text{nF}$ with $r = f_s^2/(f_p^2 - f_s^2)$. Hereby, the result $f_a \approx f_p \approx 17\,090\,\text{Hz}$ of the above discussion has been used. It can be seen that the information from the antiresonance pulls the antiresonance of the equivalent circuit into the right place and provides curves that better match the measured data all across the plotted frequency interval. The two additional cases plotted in cyan show the effect of adding and subtracting the uncertainty $\sigma = 0.4\,\text{nF}$ deduced for C_0 as a consequence of the uncertainty on the antiresonance frequencies discussed above.

P.4 Determining characteristic frequencies from Y- and Z-circles with scattered data

The impedance circle of resonator 8 with the glass bottom piston (figure O.15) is an example of how noise on the current phase data leads to scattering of the points in the complex Z plane. This means that going through the data set point by point, i. e. in clockwise direction through the circle, there are intermittent forward and backward steps. How should the characteristic frequencies (f_r,f_a,f_{mB},f_{nX},...) be determined in such a case? To take just the data point nearest to the corresponding point of the fitted circle would be suboptimal. And what should be taken as nearest, smallest distance or smallest angular offset? The approach to choose the data point with the largest G to read out f_s, the one with the largest $|Z|$ for f_n, the smallest X for f_a etc. would be even worse.

The approach followed here was to write the data set in a new coordinate system, a cylindrical one corresponding directly to the fitted Y- or Z-circle. In the cylindrical system each point is defined by an angle φ and a radius r, so in the case of an admittance circle one has

$$\begin{aligned} G &= C_G + r\cos\varphi \\ B &= C_B + r\sin\varphi \end{aligned}$$

where $C = (C_G, C_B)$ is the centre of the fitted Y-circle. Figure P.7 shows the angle φ over the frequency f for the data set underlying the Z-circle of figure O.15. Smoothing the data set can yield a function (the cyan line) which can be used for translating f into φ and vice versa. In practice this is done by linear interpolation of the smoothed data set. Calculating the characteristic frequencies that way yields much more telling values because they are not influenced by the random offset of single data points. For example, f_s can be computed by requesting the frequency value corresponding to $\varphi = 0$ from the interpolation function, or f_r upon first calculating the angle φ where the fitted circle crosses the abscissa in the G-B-coordinate system and then requesting the frequency for that angle. The lowpass filter used for the smoothing was a Butterworth filter of order two, and the cutoff has been chosen to be 0.08 times the Nyquist frequency. The filter has been applied bidirectionally. This has been done using the functions `butter` and `filtfilt` of the `scipy.signal` section of the SciPy library [223]. The entire program written for the analysis of Y- and Z-circles including this and other utility functions has been published online [432].

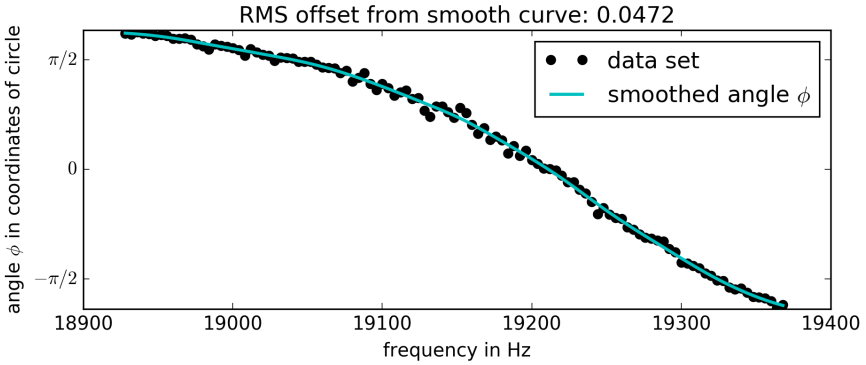


Figure P.7 Smoothing the angle function in scattered Y and Z data.
The black dots show the angle ϕ in the coordinate system of the fitted circle of each point in the raw data set which is present as a frequency sweep of equidistant steps. The cyan line shows the result of a smoothing process with a bidirectional Butterworth lowpass filter. The data set underlying the smooth curve can be used to program a function returning interpolated frequencies requested for special values of the angle φ, e. g. $\varphi = \pi/2$ for f_{mB} or f_{mX}.

P.5 Resonator Nº 8: raw characterisation data

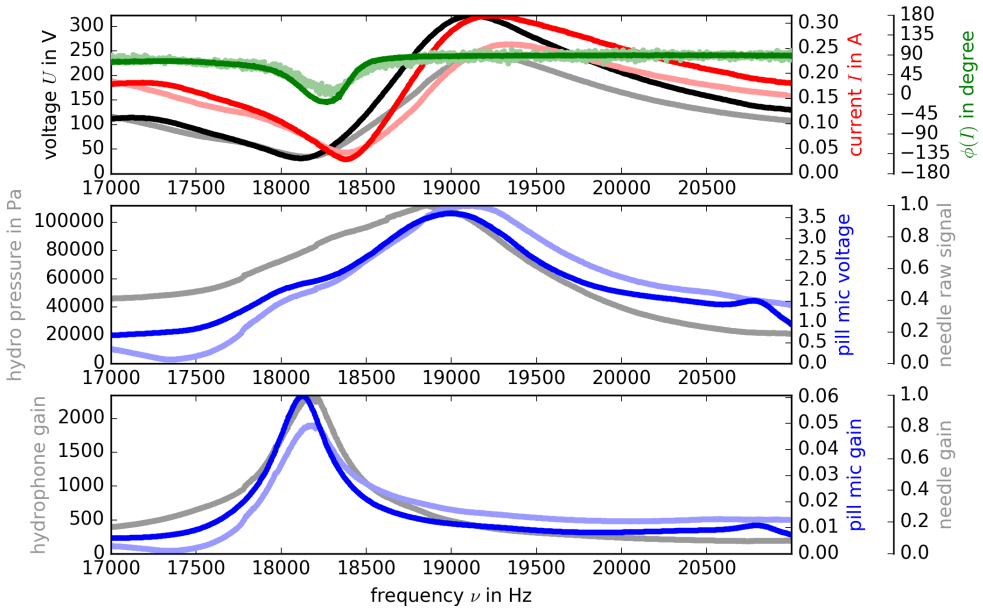


Figure P.8 Raw data of resonator Nº 8.
The diagram shows the electric raw data in the top plot and below the acoustic signals, first the amplitudes and at the bottom the gains. The fully coloured data set (case 25, representing the setup with symmetric aluminium pistons) is the basis for the Y- and X-circle analysis in appendix O.3.2, while the shaded data (case 12) has been gathered with the hydrophone tip at the centre of the resonator. The Q-factors for amplitudes and gains are 15 and 52 for the hydrophone and 18 and 52 for the mic signal in case 12. In case 25 the mic Q-factors are 19 and 66. The pressure response of case 12 is 600 $\frac{\text{Pa}}{\text{V}}$ at the pressure peak (18 856 Hz) and 2340 $\frac{\text{Pa}}{\text{V}}$ at the pressure gain peak (18 206 Hz).

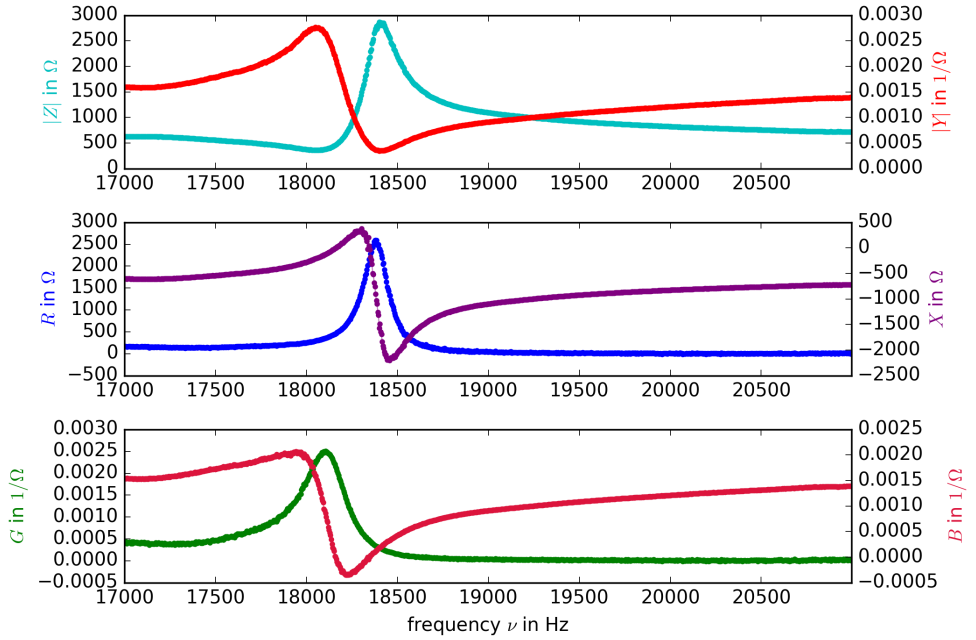


Figure P.9 Impedance and admittance of resonator Nº 8.
The data stems from case 25.

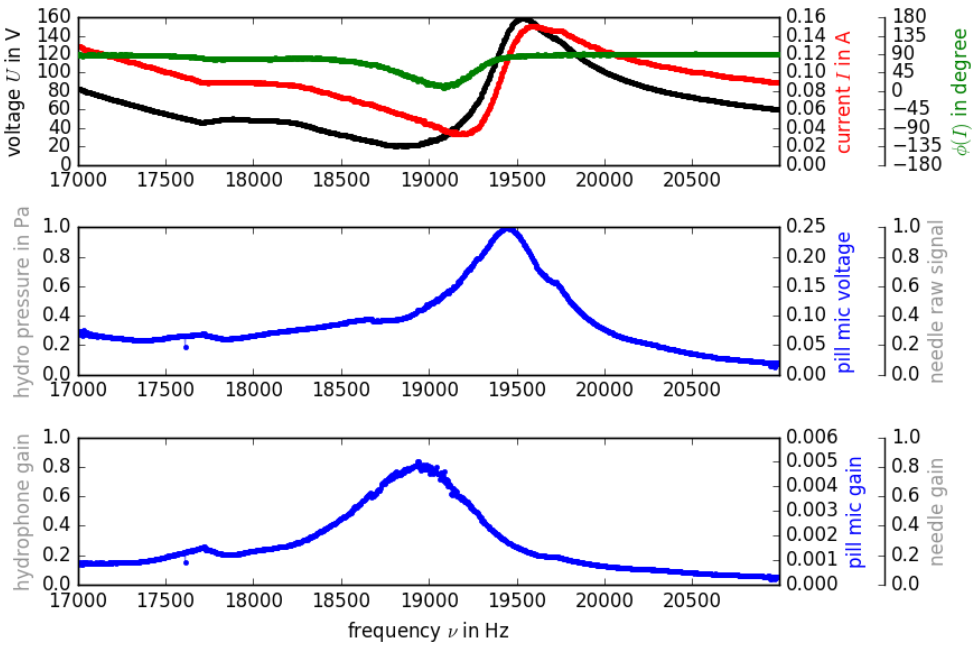


Figure P.10 Raw data of modified resonator Nº 8.
This data (case 47) has been recorded in the setup with the glass bottom piston and the standard aluminium top piston in place.

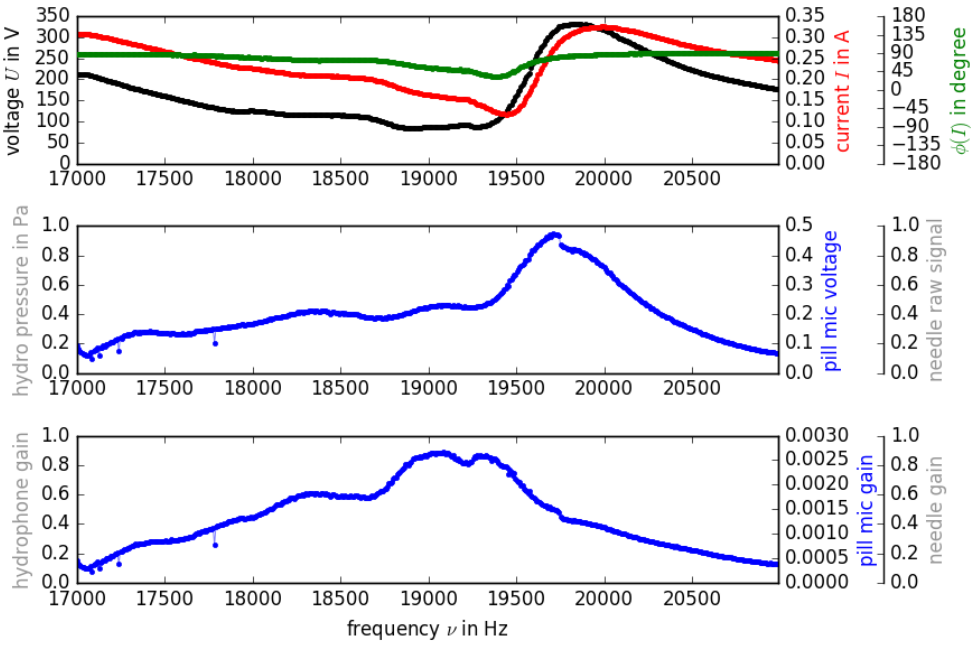


Figure P.11 Raw data of resonator Nº 8 at elevated driving voltage.
This data (case 50) shows the frequency sweep under the condition of an elevated driving voltage and has been recorded in the setup with the glass bottom piston and the standard aluminium top piston in place.

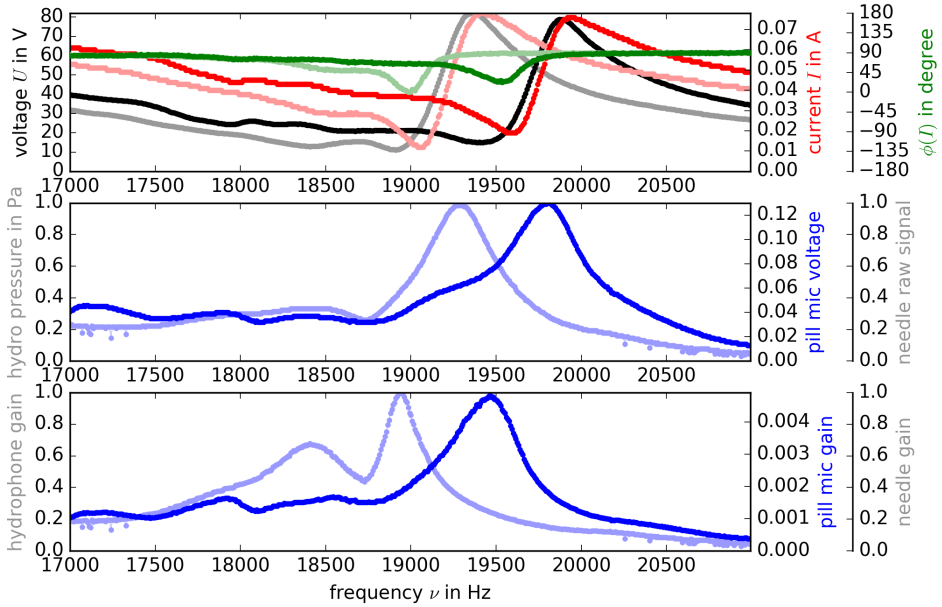


Figure P.12 Resonator Nº 8 at two different temperatures.
The speed of sound in acetone is dependent on the temperature, it grows with lower temperatures as the density and stiffness increase. This shifts the frequency of those resonances determined by the liquid. The data of case 55 (dark colours) has been recorded at −1.8 °C and the background data (case 56, shaded) corresponds to 14.1 °C. The 16 degree temperature offset shifts the resonance about 500 Hz. For the microphone signal, the problem of multiple peaks exists only at the colder temperature in this case, but similar peak-splitting has on other occasions also been observed at room temperature.

P.6 Resonator Nº 5: raw characterisation data

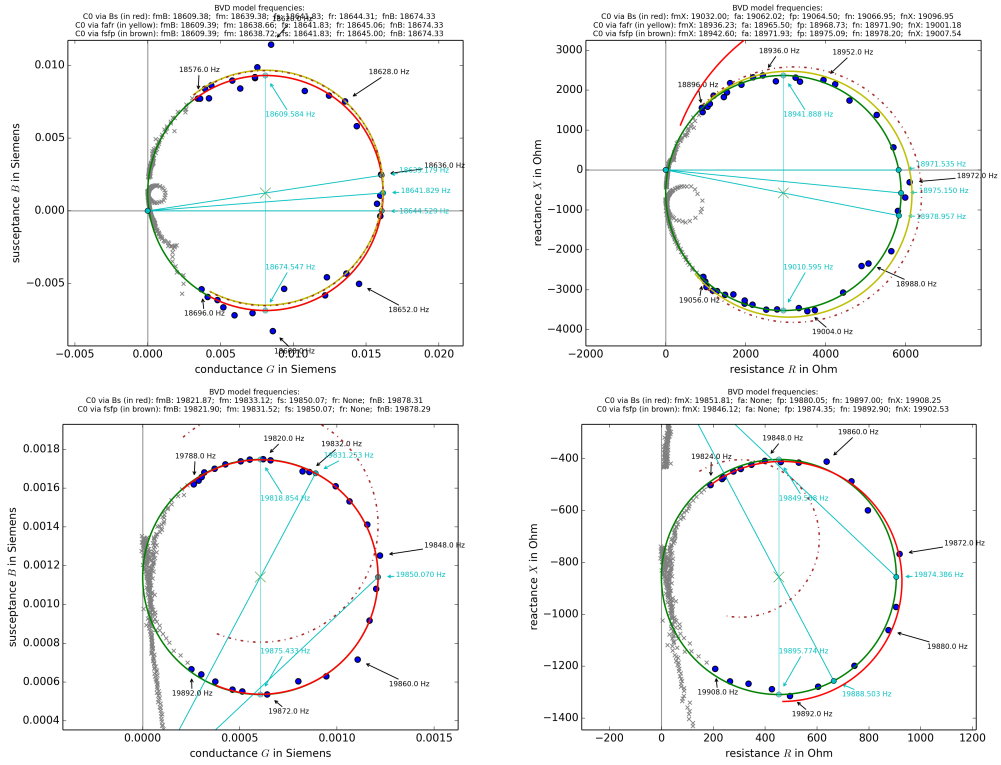


Figure P.13 Resonator Nº 5: Y- and Z-circles from setup with transformer
Y-circles are on the left, and Z-circles on the right. The top row corresponds to the first resonance, the bottom row to the second resonance. As can be seen from figure O.21 (p. 384), the transformer raises the lowest current amplitudes substantially, which leads to a better signal-to-noise ratio around the first antiresonance. This is the reason why the impedance circle in the upper right diagram is much clearer than the ones in the following figure. Unfortunately, at the time of measurements, stable driving voltage conditions were deemed to be more important than clear impedance circles, and thus most of the other characterisation data had been collected without the transformer.

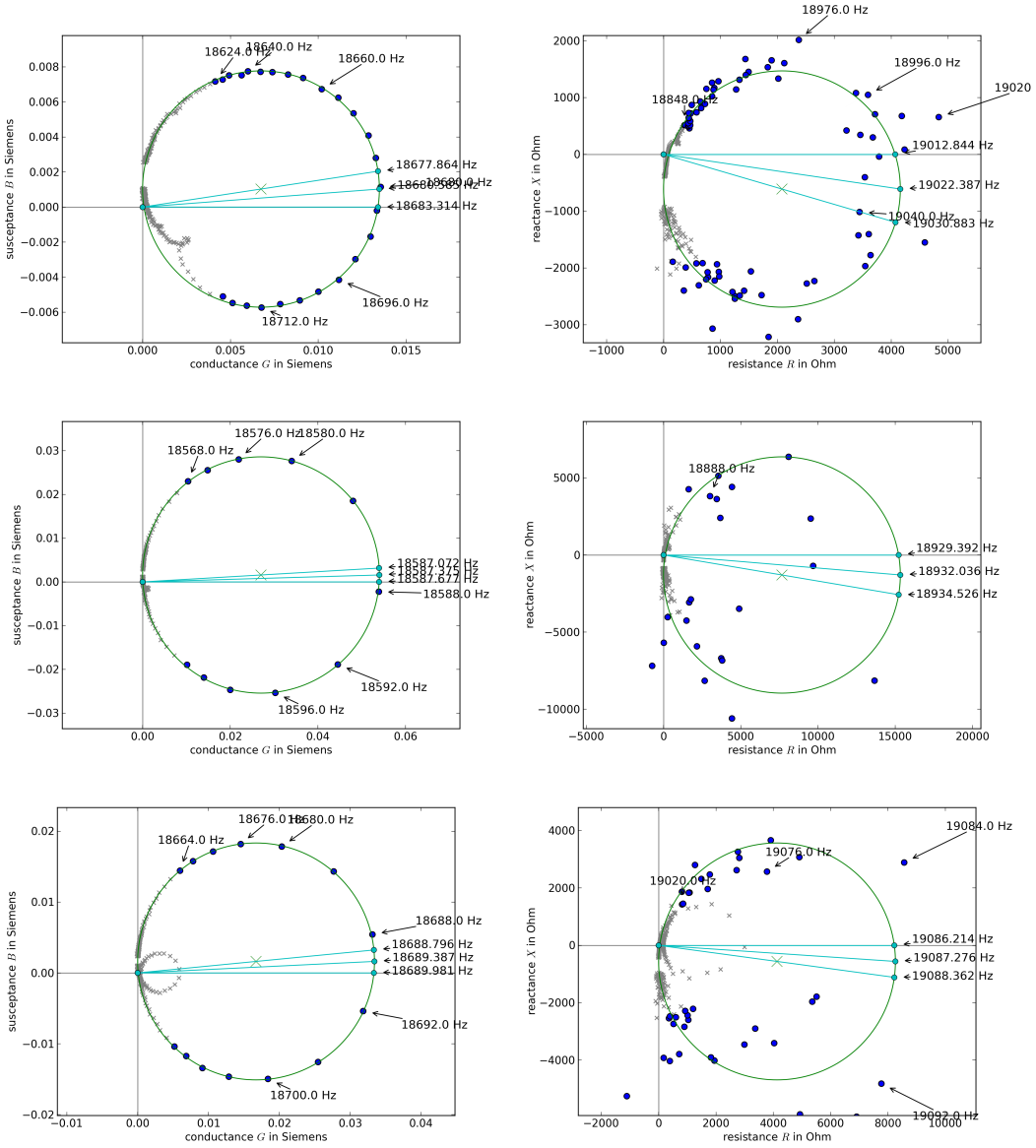


Figure P.14 Resonator № 5: Y- and Z-circles of 1st resonance.
The three datasets of cases 140, 149, and 187 have been selected for presentation because they correspond to the smallest, the largest, and an intermediate Y-circle from a series of measurements recorded with the same setup. The low quality of the scattered Z-circle data points is due to the current amplitude going to zero at the antiresonance in the setup without the transformer. The favouring of that setup was driven by the desire to keep the driving voltage constant. The plots in figure P.16 are justification for using the fitted Z-circles for the determination of characteristic frequencies of the antiresonance such as f_a and f_p. The plots on the left show that concerning admittance circles the situation is unproblematic because the data from around the resonance is much cleaner. As the discussions in appendix O.2.2 show, it is possible to fully determine a BVD equivalent circuit mainly based on the Y-circle data, i. e. based on the resonance. Only of the determination of the parallel capacitance C_0 information about the distance between resonance and antiresonance has to be utilised, and this information can be gained even with data of the quality depicted above.

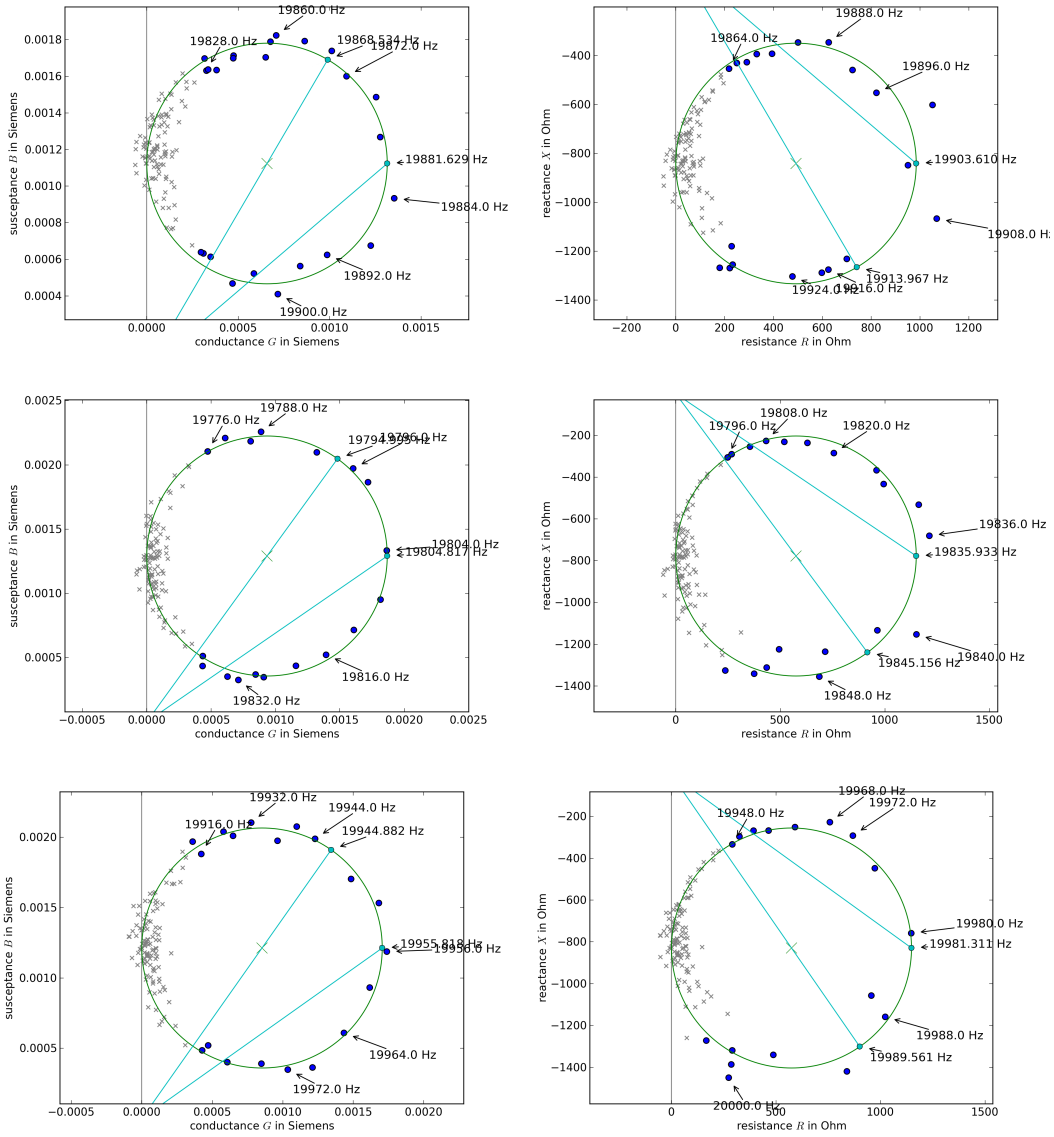


Figure P.15 Resonator Nº 5: Y- and Z-circles of 2^{nd} resonance.
In the case of the second resonance the current phase never crosses the 0° line, and the Y- and Z-circles thus do not intersect with the real axis. Consequently, no frequencies f_r and f_a can be determined. This is the reason why some formulae in table O.5 differ from table O.4.

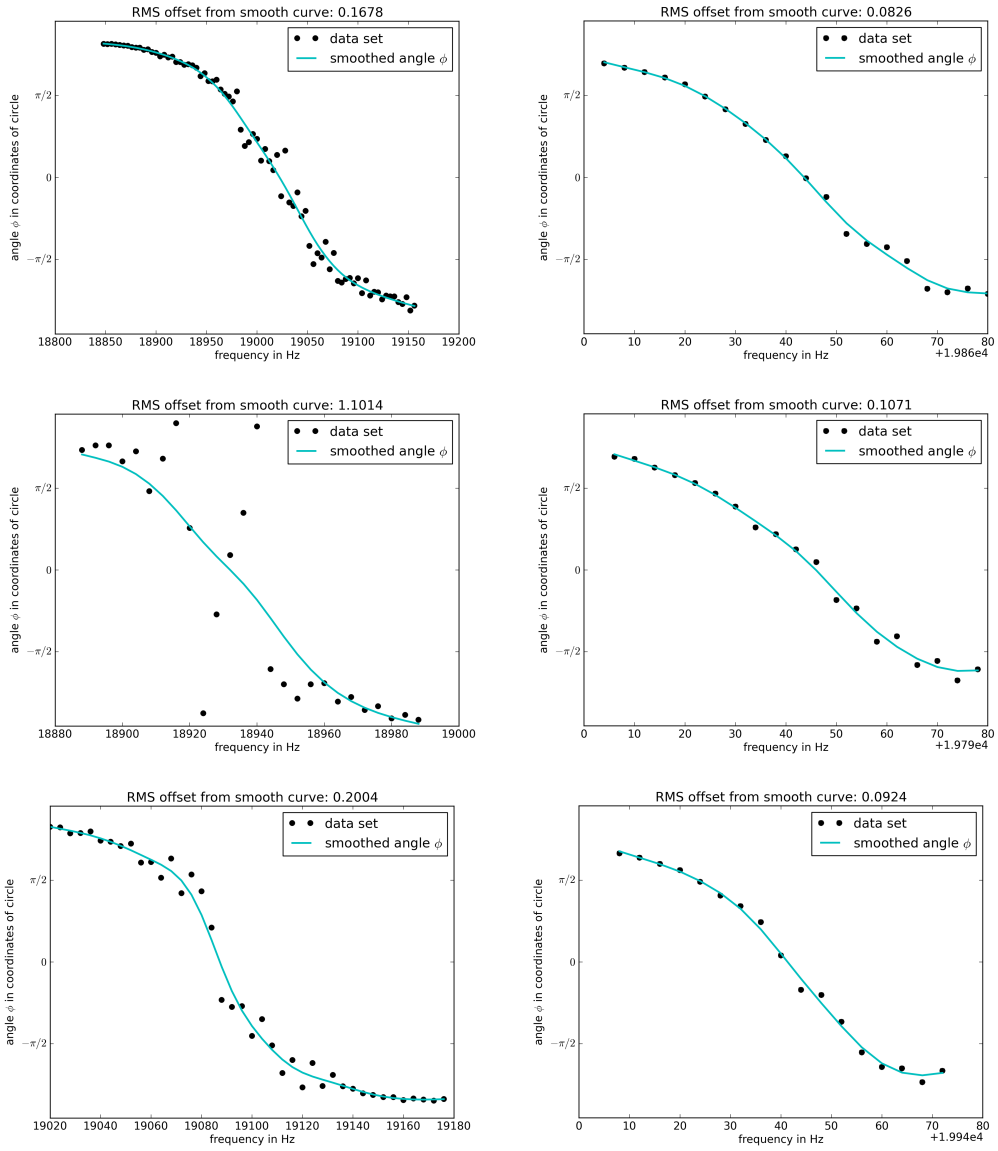


Figure P.16 Resonator Nº 5: phase angle smoothing for interpreting Z-circles. Going through the analysis formulae in tables O.4 and O.5 it can be seen that the size and position of the impedance circle is not taken into account. Only the characteristic frequencies f_a and f_p are needed. Appendix P.4 explains how a smoothing filter applied to the angle data with respect to a fitted circle allows to read out the characteristic angles even if the current-to-voltage phase data is very noisy. In these cases a look at the quality of the smoothed angle-to-frequency interpolation function is needed to check the trustworthiness of the characteristic Z-circle frequencies computed and re-used automatically in the Y-Z-circle analysis program [432]. The check plots shown above have been interpreted as positive results because (a) the angle data (black dots) is deemed to contain the structure of a smooth transition and (b) the smoothed function (cyan line) is regarded as correctly tracking that transition. The left column of plots corresponds to the impedance circles of the first resonance of cases 140, 149, and 187. The right column shows the same for the second resonance.

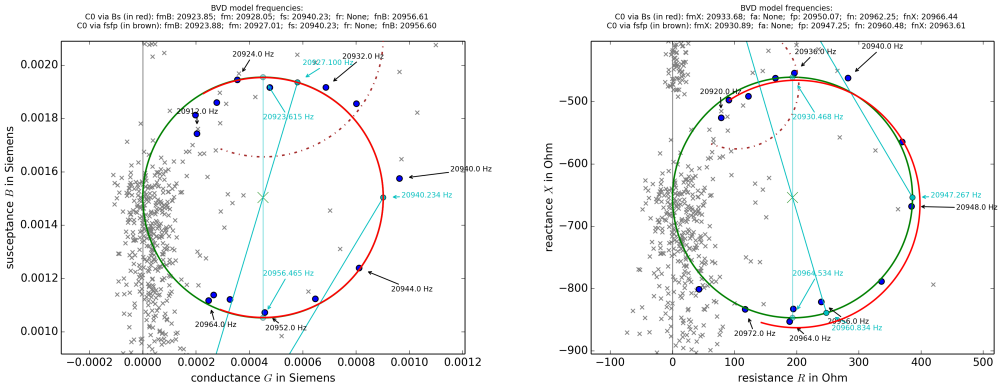


Figure P.17 Resonator Nº 5: Y- and Z-circles of 3rd resonance.
This pair of Y- and Z-circles comes from the 3rd resonance slightly below 21 kHz in the dataset of which the frequency response is plotted in figure Q.10. The corresponding BVD equivalent circuit properties are $R = 1109\,\Omega$, $L = 5.37\,\mathrm{H}$, $C = 1.08 \times 10^{-11}\,\mathrm{F}$, $C_0 = 16.0\,\mathrm{nF}$ (via f_s & f_p), $Q = 637$, $k = 0.026$.

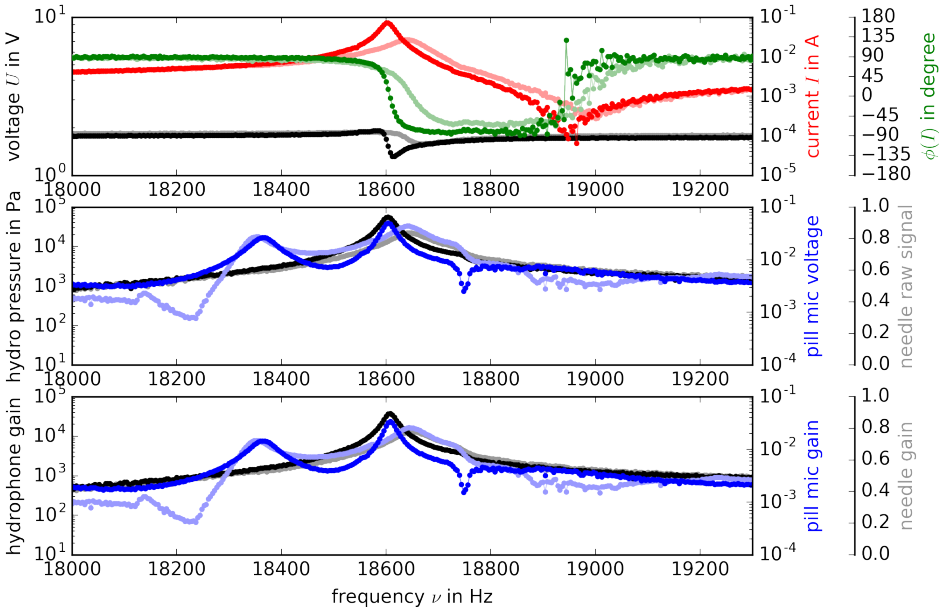


Figure P.18 Raw data comparisons with resonator Nº 5: difference in time.
A time period of eleven days lies between these two data sets (cases 144 (brightened, in background) & 145). The experimental setup has been left untouched during it, except for the refilling of 8 ml of acetone to restore the filling level at the upper edge of the 8 cm pen mark and 15 minutes of degassing with the help of sound agitation and pumping. The voltage amplitude marked in black in the top plot offers simple features for an easy comparison of this figure with the two following ones. Here, there is a difference in sharpness of the feature at 18.6 kHz, whereas in figures P.19 and P.20 it only shifts its position. The hydrophone position was 1.5 cm for all the data sets shown in these three figures.

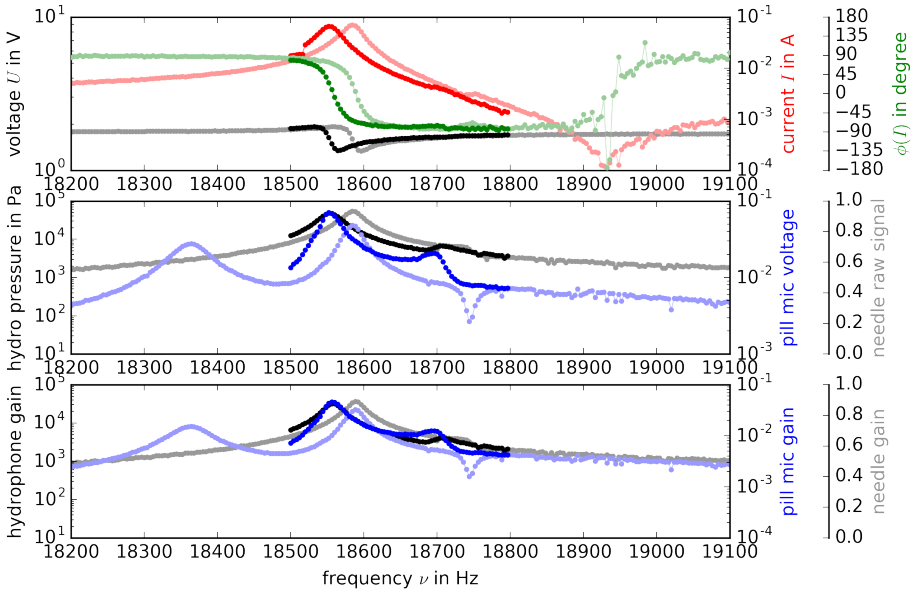


Figure P.19 Raw data comparisons with resonator № 5: difference in filling level.
These two data sets (cases 154 (brightened) & 155) were recorded before and immediately after adding 3 ml of acetone through the hydrophone outlet and degassing for a few minutes. The temperature dropped from 22.0 to 21.85 °C during degassing. Looking at the voltage again or the other signals plotted in the diagram at the top, we see that the 0.5 mm difference in acetone filling shifted the features, but did not reshape them. In the acoustic signals plotted below, the hydrophone amplitude signal also exhibits the one slightly shifting resonance, but the situation is somehow different for the microphone where a multi-peak structure not only shifts, but also changes its shape. The shape changing of that structure is assumed to be due to a changed interaction of vibration modes of the (changed) liquid volume and the (unchanged) structural materials.

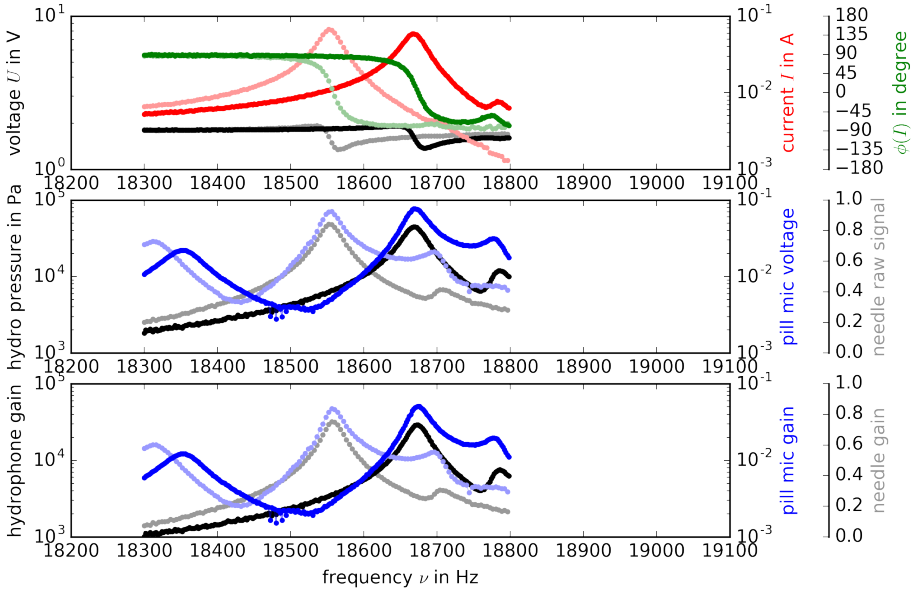


Figure P.20 Raw data comparisons w. res. № 5: difference in temperature.
These two measurements (cases 156 (brightened) & 159) were recorded for investigating the influence of a changed temperature. They are about an hour apart, during which time the chamber was cooled down from 21.85 to 18.3 °C. After finishing recording № 159 the temperature was 17.9 °C. Because resonator 5 had been setup on a normal working table instead of inside the freezer, the cooling had to occur by a small amount of snow in paper tissue placed on the aluminium top head. Therefore, some small force impact on the top head (of similar magnitude as otherwise needed for hydrophone position changes) cannot be excluded. The temperature measurement came from the thermocouple on the glass side wall and ensured that the cooling affected not only the top head, but also the liquid and through it the glass wall.

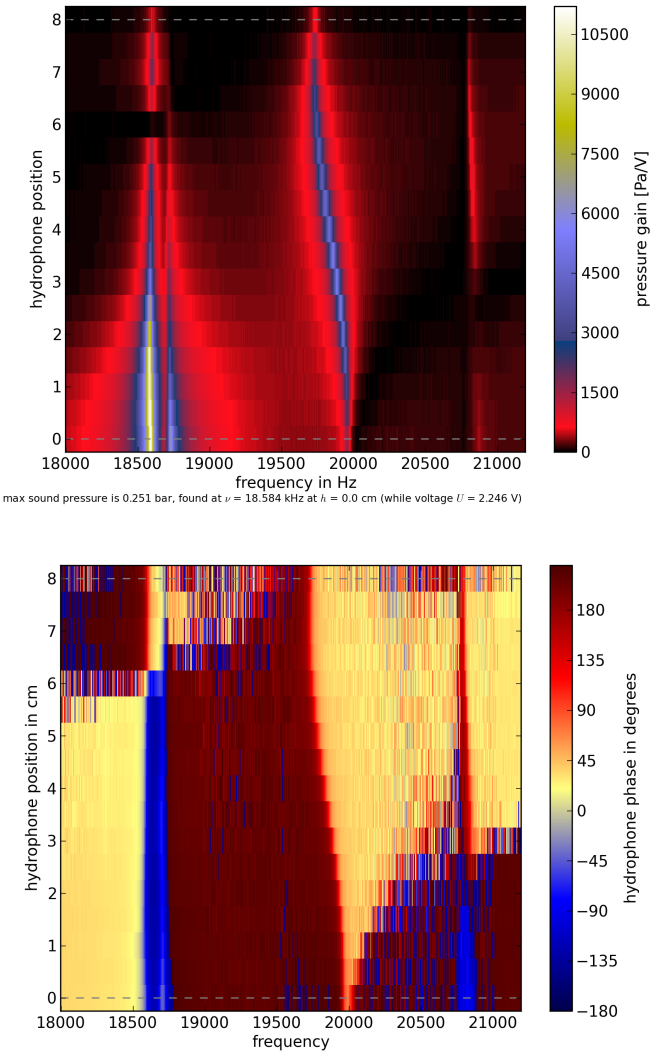


Figure P.21 Resonator Nº 5: sound pressure map recorded with later setup.
In the later setup (see figure P.2) the phase with respect to the voltage phase of each signal has been recorded as well, but the map covers not the whole liquid volume, only the vertical range possible with the hydrophone held in place by the aluminium top head is scanned. The dataset corresponds to cases 287 to 303. A notable feature and difference with respect to figure O.23 is the split in the first resonance, which may be due to the interaction of two close resonances. Figure P.23 Tries to look into more details of the sporadically appearing split.

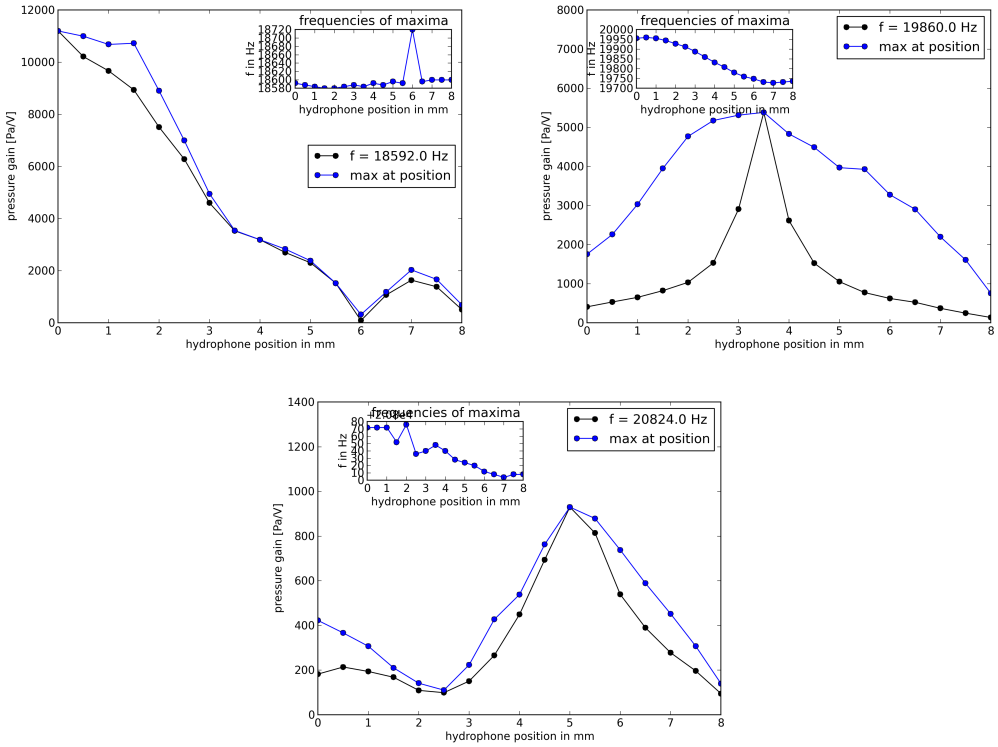


Figure P.22 Resonator № 5: sound pressure profiles recorded with later setup.
These profiles, taken from the dataset mapped in figure P.21, seem to be less affected by unstable working conditions of the resonator as the ones plotted in figure O.24, but they cover a smaller vertical distance range. Throughout the measurement campaign, from time to time the hydrophone position yielding the maximum amplitude at the second resonance has been searched manually, and values between 3.4 and 4.2 cm had been noted down. The profile above is in better agreement with those values than the one in figure O.24.

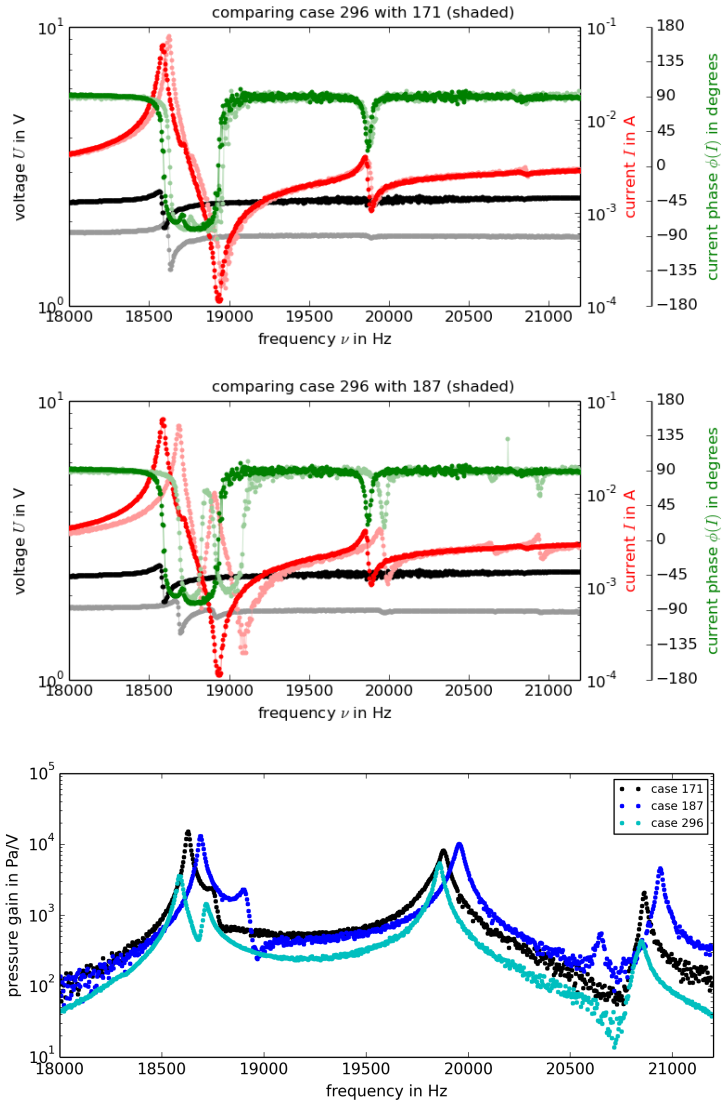


Figure P.23 Resonator Nº 5: split of the 1st resonance.
The three recordings 171, 187, and 296 are compared here, all with consistent hydrophone positions of 3.5 cm. Case 171 is part of the hydrophone map presented in figure O.23 where no split of the 1st resonance appears and only a small shoulder is visible at that height. Case 187 is a control recording with the hydrophone pulled up again to the default position directly after the last dataset of that map had been acquired. But, as the aluminium top head had been removed for the recordings below 1 cm, case 187 is the only measurement at 3.5 cm with the hydrophone hanging freely and no top head installed. Looking at the difference in the current phase of the datasets 171 and 187, one could think that the damping by the aluminium top head clamped with lightly tightened bolts against the rubber sealant ring in the groove of the lower part of the aluminium flange is what let's the split disappear. In that case the hydrophone signal of case 187 should exhibit the split. In fact it does, but only very weakly, as can be seen in the lower plot. The other problem is that dataset 296 does not fit at all into the assumed pattern. That dataset is part of the hydrophone map shown in figure P.21, where there is a strong split in the 1st resonance. But the top head had been back installed and the current phase signal has taken back the shape like in case 171. The other characteristic, in which case 187 is the exception and the other two recordings are in agreement, is the location of the peaks on the frequency axis. Normally, the temperature and the liquid volume would be considered the main parameters shifting the frequencies. Unfortunately, the influences of temperature and the top head could not be explored any more systematically during the finite measurement campaign. Testing different ways of putting the opened resonator under vacuum and fixing sound pressure probes can be deduced as suggestions for future campaigns at this point.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
a	fit parameter
B, b	fit parameters
B	susceptance
C	capacity
c	conversion factor
f	frequency; generic function
G	conductance
I	current
L	inductance
Q	quality ("pointedness" of a resonance peak)
R	resistance
r	radius; capacitance ratio
$s(t)$	time-based signal
$\hat{s}(f)$	frequency spectrum
t	time
U	voltage
X	reactance
x	generic parameter
Y	admittance
Z	impedance

List of Greek quantity symbols

Symbol	Description
α	compensation factor
μ	fit parameter
ν	frequency
σ	fit parameter
ϕ, φ	phase angle
ω	angular frequency

List of abbreviations

Abbreviation	Description
ADC	analogue-digitial conversion/converter
AI	analogue input
BF	bubble fusion
BFBM	bubble fusion Bernie & Markus (prefix for collaboratively authored routines)
BFM	bubble fusion Markus (prefix for routines of own authorship)
FFT	fast Fourier transformation
GPIB	General Purpose Interface Bus (IEEE bus specification standard)
GUI	graphical user interface

NI	National Instruments®
KIT	Karlsruhe Institute of Technology (Karlsruher Institut für Technologie)
PCB	short for picocoulomb (in the company name PCB Piezotronics®)
PCI	Peripheral Component Interconnect (local computer bus standard)
RMS	root mean square
RPI	Rensselaer Polytechnic Institute
VI	virutal instrument

Appendix Q

An FEM simulation for studying the vibration behaviour of a sonofusion resonator

In the course of the RPI-KIT collaboration project on SF it was decided to set up a new FEM simulation of the SF resonators for two reasons, (a) for better understanding the vibration behaviour of the existing resonators and being able to correctly interpret their characterisation data and (b) in order to subsequently use that understanding and the simulation capability for testing and evaluating new SF resonator design ideas. For the first purpose, several geometry setups were simulated but also several design parameter sensitivity and dependency studies were conducted. The application to the second purpose is the content of chapter 5.

Q.1 Forced harmonic analysis of a piezo-driven liquid-filled resonator

Q.1.1 The finite element model

The commercial software suite ANSYS® was chosen because it has the capabilities for the simulation of piezoelectric materials, of the force interaction between structural materials and domains of acoustic waves (i. e. fluid-structure interaction, FSI), and because of existing user experience at the IKET. The whole model from the geometry over the meshing to the solution has been scripted in APDL[1]. Short APDL code snippets containing crucial setup steps of the piezo-driven resonator simulation can be found in appendix R.

It was determined to set up a 2D axis-symmetric model as all mode shapes of interest were assumed to be of radial symmetry. Isotropic structural materials, the liquid, and the piezo crystal are of the element types `plane82`, `fluid29`, and `plane223`, respectively. Plane82 is an 8-node quadrilateral containing 4 midside nodes; there are 2 translational degrees of freedom (DOF) per node: vertical and radial displacement. Fluid29 is a 4-node quadrilateral with 2 translational DOF and

[1] short for *Ansys parametric design language*, the log and scripting language of Ansys Classic

a 3rd DOF, pressure. Plane223 is an 8-node quadrilateral multi-purpose coupled-field element capable of up to 4 DOF per node. The latter, in its piezoelectric option, results in 3 DOF per node, two accounting for translation and one for voltage. The adopted reaction forces corresponding to the three DOF are then 2 times force and electrical charge. Plane82 allows for Rayleigh damping whereby the damping matrix is a linear combination of the mass matrix and the stiffness matrix and a frequency-independent damping ratio or loss tangent (see equations K.45 & K.48). Plane223 can on top account for a dielectric loss tangent, whereas `fluid29` allows no viscous damping inside the fluid domain. The coloured mesh of an idealised resonator geometry in figure Q.1 indicates how those element types are used.

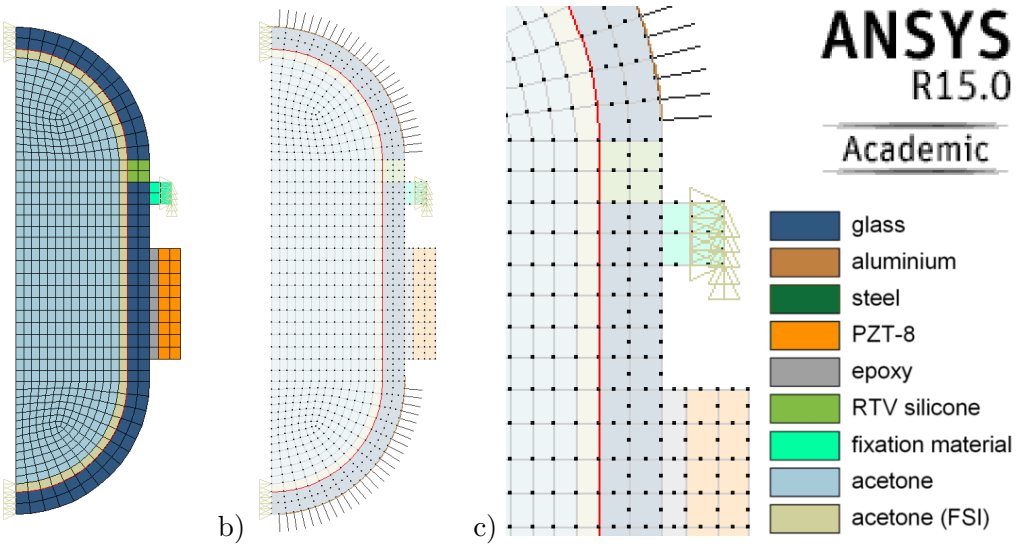


Figure Q.1 FE mesh of an idealised resonator geometry.
The examined resonators are vessels filled with liquid, and glued to the outside of the vessel is a piece of piezoelectric material serving as the actuator exciting a vibration motion of the whole compound. The part of the mesh representing the transducer made of piezoelectric material is depicted in orange and is made of the element type `plane223`. It is coupled by the epoxy glue (grey) to the vessel's glass wall (dark blue, like all other structural parts modelled with `plane82` elements). The vibration motion of the glass wall is directly coupled to the pressure and motion waves in the liquid, and the FE model has to account for that. This problem is solved by the `fluid29` elements which can exist in two versions, with or without the translational nodal DOF activated. Throughout the acoustic domain (light blue) the displacement DOF are turned off. Only for those `fluid29` elements in contact with structural elements (depicted in beige) they are activated. In the two node plots the faces between the nodes shared by the fluid and the structural elements are marked by the thin red line showing that the FSI routines apply there. These nodes can move, and within the equations of the outer row of `fluid29` elements the displacement of these element faces is translated into pressure contributions and vice versa, as outlined in appendix chapter K.3. The green elements represent one more structural material, silicone, used to connect assembly parts in a softer way than epoxy does. The cyan area represents an unnaturally soft and light material used for a low-impact anchoring in space and preventing rigid body modes. Some BC settings are symbolised in the mesh by grey triangles: zero radial displacement for structural nodes on the central axis and no displacement at all for the outer rim of the ring of fixation material. The voltage BCs on the transducer electrodes are not shown. What appears as needles on the rounded top and bottom head surfaces in the node plot are special rotated nodal coordinate systems that were defined for this subset of nodes for allowing the output of the normal and in-plane components of the displacement.

The radially polarised piezoelectric transducer is shown in the figure in orange. Its inside and outside surfaces are the electrodes (as in the original where these surfaces are silver-coated). The excitation of the vibration motion occurs through the application of a harmonic voltage signal to one of these electrodes while the

other is kept at ground potential. It is done by setting the voltage DOF of all nodes belonging to one electrode to zero and applying a constant nonzero voltage $U_0 = 100\,\mathrm{V}$ as the boundary condition (BC) to the other electrode. In the framework of the harmonic analysis this mirrors the forcing through a sinusoidal voltage signal of amplitude U_0.

Another important coupling mechanism is the one between the vibration motion of the structural parts and the sound pressure field in the acoustic domain (fluid-structure interaction, FSI). FSI is accomplished by the outer layer of `fluid29` elements shown in beige in figure Q.1. While throughout the acoustic domain each node has just one single DOF, pressure, the nodes shared by `fluid29` and `plane82` elements have three DOF for displacements and pressure. Force loads are transferred between the two domains over these nodes. The conversion of sound pressure into force and back in static, transient, and harmonic oscillation cases is accomplished within the `fluid29` elements as outlined in appendix chapter K.3 or explained in more detail in [8].

The fixation in space of the resonator models occurs through defining a zero displacement BC for a subset of nodes. In figure Q.1 that boundary condition is marked by triangles. It is the material shown in light cyan that connects the resonator to the fixation nodes. This material was associated with the very low values for density and Young's modulus listed in table Q.3 resulting in a very soft fixation of the resonator. Instead of a rigid fixation type it has been deemed better to keep the impact of the fixation on the vibration eigenmodes of the resonator as low as possible both in the model and in the lab experiment where the resonators had been suspended by wires.

The air environment surrounding the resonators (sound emission into this domain and losses due to it) was generally not modelled. The entire acoustic fluid domain and the FSI mechanism were also assumed lossless. The only loss mechanisms applied were dielectric and structural damping. For the introduction of dielectric damping in the piezoelectric material Ansys offers the command `mp,lsst` for specifying a dielectric loss tangent, meaning that an imaginary part is added to the dielectric permittivity. For energy dissipation in the bending structural materials Ansys offers various ways of composing a damping matrix as visible from equation K.48. The coefficients ζ and ζ_i in that equation are used to set frequency-independent damping ratios. Only these two options were used. The coefficient ζ can be set by the `dmprat,r` command resulting in a uniform damping ratio `r` applied to all structural elements of a model. The command `mp,dmpr,i,r` allows to set the material-specific coefficients ζ_i. The loss mechanism arising from AC current leaking through a parallel capacitance C_0 of the transducer (see appendix J) can be taken into account by creating a `circu94` element for simulating the capacitance and connecting[2] the two electrodes through it.

The multi-purpose coupled-field elements `plane223` offer quite some implemen-

[2]In the discussed forced harmonic analyses the voltage DOF of the nodes forming the electrodes are fully determined by the voltage BCs. This means, any additional coupling of the voltage DOF within electrode surfaces would have no effect. But in the case of connecting `circu94` elements to the electrodes, a particular form of DOF coupling is necessary, as it allows the `circu94` to access the sum of nodal charges through a *master node* of the coupled set [7].

tation flexibility. Their setup is therefore not trivial. The example VM-231 of the Ansys verification manual [9] was taken as the reference for creating the APDL code used in most simulations. The resulting APDL code is listed in appendix R.1. The verification manual cases VM-175, VM-176, and VM-237 offer alternative reference codes. Only in those validation cases listed in table Q.4 where the piezoelectric strain tensor entries were used for the APDL input instead of the piezoelectric stress tensor entries the alternative input example VM-237 was adapted. The internet provides some additional and potentially useful documentation[3] on the topic of setting up piezoelectric materials in Ansys.

Q.1.2 Material properties

Material constants of the piezoelectric ceramic

PZT is the common[4] abbreviation of *lead-zirconate-titanate* which is a piezoelectric material normally manufactured in polycrystalline form as ceramic. The chemical formula can be written as $Pb(Zr,Ti)O_3$, meaning that zirconium and titanium atoms as group 4 elements can replace each other in the crystal lattice. Hence, a different way of writing the formula is $PbZr_xTi_{1-x}O_3$. Some particular mixtures of this class which are more commonly used have extra labels, like PZT-8 which stands for a material with $x = 0.52$ [263]. PZT-8 has properties in accordance with the *Navy type III* classification of sonar transducer materials which means that dielectric losses and heating are minimised in order to be suitable for high-power and high-voltage applications [352]. This goes together with type III materials being not easily depolarised, thus one speaks of a *hard* piezoelectric material [516]. Commercial suppliers have extra names for their materials. The examined transducers were manufactured by Channel Industries®, Santa Barbara, CA, who call their two PZT-8-type materials C5800 and C5804. The PZT hollow rings driving the here discussed RPI SF resonators were made of C5800. Table Q.1 lists various sets of material properties for the relevant piezoceramics, gives literature references, and indicates which ones were used for the presented FEM simulations.

Liquids

The liquid of interest here is acetone, its acoustic properties are listed in table Q.2. The acoustic properties of the fluid determines the emerging sound field patterns in a resonator and influences the vibration response of the resonator's structural parts. The reasons for considering acetone as a working fluid for SF experiments are outlined in chapter 1.4.1 and in more detail in [326]. Water is also listed because a

[3] These resources may be useful: the APDL macro `piezmat.mac`, to be found e. g. at `http://web.mit.edu/mkt/Public/ANSYS/piezmat.mac`, the memo on piezoelectric material data input by Sheldon Imaoka (`http://ansys.net/tips/Week13_TNT_Conversion_of_Piezoelectric_Material_Data.pdf`), and the Excel® macro `Conversion_of_Piezoelectric_Material_for_Ansys.xls` and its documentation at `http://smartmaterials.free.fr/ressources.html`. A blog entry by Mohamed Senousy at `http://www.ansys-blog.com` of October 2013 on APDL-less definition of piezoelectric materials in Ansys Workbench (version 14.5 and later) might also be noteworthy.

[4] although the term "PZT" was trademarked by Vernitron Inc., Cleveland, OH [2]

Table Q.1 Material properties of PZT ceramics

The material constants are listed for three types of PZT ceramics according to multiple literature sources. The references are given at the bottom of the table. The material PZT-4 is included because it was used in the code-to-code comparison with the simulations of Cancelos [69] for validation purposes. For the specification of piezoelectric materials in Ansys according to example VM-231 [9] the whole set of nonzero stiffness matrix entries c_{ij}^E is needed, and therefore, the first of the two sets with the short label "p4a" was used. Table Q.4 shows that it is a consistent dataset, which means that either using the piezoelectric stress tensor $\boldsymbol{e}$ or the piezoelectric strain tensor $\boldsymbol{d}$ yields the same results. While PZT-4 is a Navy type I material, all later SF resonators manufactured at RPI have transducers made of Navy-III piezoceramics, namely the C5800 ceramic of Channel Industries. PZT-8 is included in the table because it is often the Navy-III material of reference in literature. Secondly, as the set "C58a" does not readily list all stiffness tensor entries, the PZT-8 literature values together with the C5800 specifications in [353] and additional data from Cancelos were the basis of a preliminary PZT-8 dataset (short label "p8d"). A large part of the FEM calculations presented here were conducted with these properties, in particular the parameter study and the EA optimisation runs. With the availability of the second C5800 dataset ("C58b"), which is consistent too, all validation-related FEM computations were redone.

property	factor	unit	PZT-4	PZT-4	PZT-8	PZT-8	PZT-8	PZT-8	C5800	C5800
ϱ		$\frac{\mathrm{kg}}{\mathrm{m}^3}$	7500	7500	7600	7600	7600	7550	7550	7550
c_{11}^E	10^{10}	$\frac{\mathrm{N}}{\mathrm{m}^2}$	13.9	13.9	–	13.7	13.7	13.7	–	17.14
c_{12}^E	10^{10}	$\frac{\mathrm{N}}{\mathrm{m}^2}$	7.78	–	–	–	–	6.97	–	10.38
c_{13}^E	10^{10}	$\frac{\mathrm{N}}{\mathrm{m}^2}$	7.43	–	–	–	–	7.16	–	9.61
c_{33}^E	10^{10}	$\frac{\mathrm{N}}{\mathrm{m}^2}$	11.5	11.5	–	12.3	12.3	12.35	–	13.00
c_{44}^E	10^{10}	$\frac{\mathrm{N}}{\mathrm{m}^2}$	2.56	–	–	–	–	3.14	2.9	2.90
c_{66}^E	10^{10}	$\frac{\mathrm{N}}{\mathrm{m}^2}$	3.06	–	–	–	–	3.37	–	3.38
ϵ_1^σ			1475	1475	–	1290	1290	1290	1400	1400
ϵ_3^σ			1300	1300	1000	1000	1000	1000	1100	1100
ϵ_1^ε			730	–	–	?	900	–	–	–
ϵ_3^ε			635	–	600	?	580	–	–	–
e_{15}		$\frac{\mathrm{N}}{\mathrm{V\,m}}=\frac{\mathrm{C}}{\mathrm{m}^2}$	12.7	12.7	–	10.4	10.4	10.4	–	11.37
e_{31}		$\frac{\mathrm{N}}{\mathrm{V\,m}}=\frac{\mathrm{C}}{\mathrm{m}^2}$	-5.2	-5.2	–	-4.0	-4.0	-4.0	–	-3.94
e_{33}		$\frac{\mathrm{N}}{\mathrm{V\,m}}=\frac{\mathrm{C}}{\mathrm{m}^2}$	15.1	15.1	–	13.2	13.2	13.2	–	13.66
d_{15}	10^{-12}	$\frac{\mathrm{m}}{\mathrm{V}}=\frac{\mathrm{C}}{\mathrm{N}}$	496	496	–	330	330	–	390	392.07
d_{31}	10^{-12}	$\frac{\mathrm{m}}{\mathrm{V}}=\frac{\mathrm{C}}{\mathrm{N}}$	-123	-123	-93	-97	-97	–	-107	-105.22
d_{33}	10^{-12}	$\frac{\mathrm{m}}{\mathrm{V}}=\frac{\mathrm{C}}{\mathrm{N}}$	289	289	218	225	225	–	245	260.55
g_{15}	10^{-3}	$\frac{\mathrm{V\,m}}{\mathrm{N}}=\frac{\mathrm{m}^2}{\mathrm{C}}$	39.4	38.0	–	29.0	29.0	–	31.5	31.66
g_{31}	10^{-3}	$\frac{\mathrm{V\,m}}{\mathrm{N}}=\frac{\mathrm{m}^2}{\mathrm{C}}$	-11.1	-10.7	-10.5	-10.9	-10.9	–	-11.0	-10.80
g_{33}	10^{-3}	$\frac{\mathrm{V\,m}}{\mathrm{N}}=\frac{\mathrm{m}^2}{\mathrm{C}}$	26.1	25.1	24.5	25.4	25.4	–	25.2	26.75
h_{15}	10^8	$\frac{\mathrm{V}}{\mathrm{m}}$	19.7	–	–	–	–	–	–	14.35
h_{31}	10^8	$\frac{\mathrm{V}}{\mathrm{m}}$	-9.2	–	–	–	–	–	–	-7.49
h_{33}	10^8	$\frac{\mathrm{V}}{\mathrm{m}}$	26.8	–	–	–	–	–	–	28.49
T_{Curie}		°C	328	328	300	?	300	–	>300	>300
Q_{mech}			500	500	1000	?	1000	1000	1100	1100
$\tan\delta_{\mathrm{mech}}$					–	?		0.001		–
$\tan\delta_{\mathrm{diel}}$	at low field		–	–	–	?	–	0.004	0.004	0.004
$\tan\delta_{\mathrm{diel}}$	at 2 $\frac{\mathrm{kV}}{\mathrm{cm}}$		0.02	–	–	?	–	–	0.007	0.007
$\tan\delta_{\mathrm{diel}}$	at 4 $\frac{\mathrm{kV}}{\mathrm{cm}}$		0.04	–	0.01	?	–	–	0.01	0.01
reference			[516]	[2]	[516]	[256]	[2]	[70, 353, 516]	[353]	[353, 359]
short label			p4a	p4b	p8a	p8b	p8c	p8d	C58a	C58b

Table Q.2 Acoustic properties of the liquids.
The listed properties are density ϱ, speed of sound c at 25 °C, and the change of the latter under temperature variation, $\frac{\partial c}{\partial T}$.

	ϱ [g/cm^3]	c [m/s]	$\frac{\partial c}{\partial T}$ [m/(s K)]
acetone [85, 401]	791	1174	-4.5
water [85]	1000	1496.7	2.4
water [69]	1000	1490	2.4

water-filled resonator was simulated in the context of validating the FE simulations by comparison with the work of Cancelos [69].

Isotropic structural materials

The so far existing SF resonators in connection with this project are all made of glass due to its elastic properties, the low damping, its easy workability, but also the possibility to see bubble clusters or sonoluminescence flashes through it. The other materials used in those designs involved silicone rubbers for connecting glass parts and epoxy for glueing the transducer to the glass. With the introduction of the flanges aluminium was added to the structural materials. The material properties of a common steel were used to model the hydrophone. Table Q.3 lists the densities, Young's moduli, Poisson ratios of these materials and indicates the literature references.

Concerning damping ratios (mechanical loss tangents $\eta = E''/E' = \tan\delta = 1/Q$), the following values were gathered from literature:

- **glass:** Zhang [535] reports a loss factor of $\eta = 0.5 \times 10^{-3}$ at $f = 8000\,\mathrm{Hz}$
- **aluminium:** $\eta = 1 \times 10^{-4}$ according to [258].
- **epoxy:** According to Chung [87], the loss factor of epoxies is 0.03. In that context, Wang et al. [499] report on the strong temperature- and frequency-dependence of the loss tangent of epoxies near the glass transition temperature T_g, and that this sensitive temperature range as a function of chemical composition can be made to overlap with ambient conditions.
- **silicone:** $\tan\delta = 0.005$ according to Sid [420] or 0.05-0.1 according to Santawisuk [398].

For polymers like silicone and epoxy a complication consists in the dependency of material properties on curing conditions and ageing. On top, the degree of damping of a vibrating structure is not only determined by the material properties but also by geometric form and mode shape. Thus, the most sensitive material constants should ideally be deduced from calibration against experimental observations generated from samples under comparable loading conditions and having comparable sizes, shapes, and preparation histories.

Table Q.3 Properties of the structural materials.
The listed properties are density ϱ, Young's modulus E, and Poisson ratio ν. The material listed in the last line, the "fixation material" with the Young's modulus and the density of silicone divided by hundred is used in the FE models as a low-impact way of fixing the meshes in space.

	E [N/m^2]	ϱ [kg/m^3]	ν
glass-1 [404]	63×10^9	2230	0.20
glass-2 [118]	64×10^9	2230	0.20
glass-3 [69]	73×10^9	2540	0.22
steel [497]	193×10^9	8000	0.28
epoxy-1 [69]	5.86×10^9	1190	0.3
epoxy-2 [497]	10^9	1215	0.3
silicone [497]	90×10^6	1040	0.485
fix. mat.	0.9×10^6	10.4	0.485

Q.2 Validation

Q.2.1 Simple static load cases

The piezoelectric ceramic was modelled in Ansys with the help of `plane223` elements and their coupled field capabilities. The corresponding lines of APDL code are given in appendix R.1. In order to verify the correctness of the resulting finite element implementation, four static load cases of simple geometries were examined. They were chosen from a list of exemplary verification cases suggested in a catalogue of Channel Industries [353] who supplied the transducer rings. The first one is the case of a thin rod or fibre of piezoelectric material where both the material's polarisation axis and the external electric field are pointing in the fibre direction. In the limit case of infinite rod length and infinitesimal diameter the analytical result for the length change of a piece of unit length can be written in a simple form involving only one of the piezoelectric constants d_{ij} and the applied voltage U. Table Q.4 lists the formulae and results for this and three other verification cases. The examined materials were PZT-4, the preliminary set of PZT-8 constants, and C5800 by Channeltech. By the match ratios given in the last column it can be seen, that unlike the other two sets, the preliminary PZT-8 data is not consistent.

Table Q.4 PZT FE model verification under static load.

This table lists deformation values in a comparison of analytical results with FE models. This serves as a check for verifying the correct implementation and setup of the piezoelectric finite elements. There are four cases representing the three material choices "PZT-4", "PZT-8", and "C5800", and for the latter material two alternative input modes for the piezoelectric matrix, either specifying the piezoelectric stress coefficients e_{ij} or the strain coefficients d_{ij}. The geometries are: rod, plane, and hollow cylinder. The deformation modes are the length of the rod, the width of a long thin plate, the thickness shear of a thin plate, and the contraction/expansion of a hollow cylinder. The rod dimensions are length l and diameter d; the plate dimensions are length l, width w, and thickness t; for the thin wall hollow cylinder it is length l, wall thickness t, and the outer and mean diameters d_o and d_m. The third column lists the conditions for approaching the limit case under which the analytical result, the formula listed in the fourth column, becomes valid. The next column shows the dimensions of the geometry covered by the FE mesh. In all these cases it is regular 2D meshes made of quadrilaterals covering rectangular areas. While in the case of the rod the 2D geometry is axis-symmetric and the mesh extends also along the long side $l \gg d$, in the cases with plates the dimension of the long edge is orthogonal to the drawing plane and thus $l = \infty$. The cross section of the hollow cylinder is a similar rectangular mesh. The last three columns list the displacement values in nanometres from the calculation, the simulation, and their ratio. Since the analytic formulae are based on the d_{ij}, it follows that where the e_{ij} were used for the FEM calculation a ratio close to 1 indicates that the set of material constants is self-consistent, and where the d_{ij} were used also on the FEM input side it only indicates that the coefficients were entered correctly.

description	material	e/d	conditions	formula	model dim. [mm]	mesh	target [nm]	FE [nm]	ratio
thin rod, parallel longitudinal mode	PZT-4	e	$l > 3d$	$\Delta l = d_{33} U$	10×200	3×100	0.28900000	0.29039995	1.00484413
thin rod, parallel longitudinal mode	PZT-8	e	$l > 3d$	$\Delta l = d_{33} U$	10×200	3×100	0.24500000	0.21549064	0.87955364
thin rod, parallel longitudinal mode	C5800	e	$l > 3d$	$\Delta l = d_{33} U$	10×200	3×100	0.26054500	0.25967264	0.99665177
thin rod, parallel longitudinal mode	C5800	d	$l > 3d$	$\Delta l = d_{33} U$	10×200	3×100	0.26054500	0.25967264	0.99665179
thin plate, transverse width mode	PZT-4	e	$w < l/3$; $w > 3t$	$\Delta w = \frac{d_{31} w}{t} U$	1×5	3×13	0.61500000	0.61907981	1.00663384
thin plate, transverse width mode	PZT-8	e	$w < l/3$; $w > 3t$	$\Delta w = \frac{d_{31} w}{t} U$	1×5	3×13	0.53500000	0.47109104	0.88054400
thin plate, transverse width mode	C5800	e	$w < l/3$; $w > 3t$	$\Delta w = \frac{d_{31} w}{t} U$	1×5	3×13	0.52607500	0.52607556	1.00000107
thin plate, transverse width mode	C5800	d	$w < l/3$; $w > 3t$	$\Delta w = \frac{d_{31} w}{t} U$	1×5	3×13	0.52607500	0.52607500	1.00000000
plate, thickness shear	PZT-4	e	$t < w/5$; $t < l/5$	$\Delta x = d_{15} U$	1×5	3×13	0.49600000	0.49609375	1.00018901
plate, thickness shear	PZT-8	e	$t < w/5$; $t < l/5$	$\Delta x = d_{15} U$	1×5	3×13	0.39000000	0.33121019	0.84925690
plate, thickness shear	C5800	e	$t < w/5$; $t < l/5$	$\Delta x = d_{15} U$	1×5	3×13	0.39207100	0.39207142	1.00000108
plate, thickness shear	C5800	d	$t < w/5$; $t < l/5$	$\Delta x = d_{15} U$	1×5	3×13	0.39207100	0.39207100	1.00000000
thin hollow wall cylinder	PZT-4	e	$d_o > 8t$; $l > t/2$	$\Delta d_m = \frac{d_{31} d_m}{t} U$	3×25, $d_m = 65$	4×34	1.33250000	1.33576328	1.00244899
thin hollow wall cylinder	PZT-8	e	$d_o > 8t$; $l > t/2$	$\Delta d_m = \frac{d_{31} d_m}{t} U$	3×25, $d_m = 65$	4×34	1.15916667	1.01582639	0.87634196
thin hollow wall cylinder	C5800	e	$d_o > 8t$; $l > t/2$	$\Delta d_m = \frac{d_{31} d_m}{t} U$	3×25, $d_m = 65$	4×34	1.13982917	1.13495533	0.99572406
thin hollow wall cylinder	C5800	d	$d_o > 8t$; $l > t/2$	$\Delta d_m = \frac{d_{31} d_m}{t} U$	3×25, $d_m = 65$	4×34	1.13982917	1.13495411	0.99572300

Q.2.2 Frequency response of the free transducer

The transducers used for all discussed RPI SF resonators are thin-walled hollow cylinders made of C5800 polarised in the radial direction and having the silver electrodes on the cylindrical inside and outside surfaces. The hollow cylinders have a wall thickness of 3 mm and a height of 25 mm. The inside and outside diameters are 65 and 71 mm. Being interested not in torsional modes but only in radial oscillation movements, the transducer can be modelled by a simple two-dimensional rectangular FE mesh in an axisymmetric coordinate system which covers the cross section of the PZT ring. The 3 mm × 25 mm area was meshed with 4×32 elements. As boundary conditions for the voltage degree of freedom the inner electrode is grounded while the nodes forming the outer one are associated with a constant value of 100 V. In the framework of the harmonic analysis type this setting corresponds to a sinusoidal signal with an amplitude of 100 V serving as the driving force. As concerns the translational DOF, there is no need for a fixation in space along the radial coordinate r in an axisymmetric coordinate system (r, z). For fixation along the z-axis, displacements in that direction were forced to zero for all nodes in the plane orthogonal to the z-axis and going through the transducer's centre of mass.[5]

In the frequency domain the FE model can be validated through two comparisons: a comparison of the simulated resonance peak with the resonance frequency prediction based on tabulated frequency constants and a comparison with the lab measurement of the transducer's frequency response. Both comparisons are shown in figure Q.2.

The applicable frequency constant for the radial expansion and contraction mode is the constant N_c given by the supplier's catalog [353] as $N_c = 42\,\text{kHz in} = 1070\,\text{Hz m}$, but an updated value of $N_c = 43\,\text{kHz in} \approx 1095\,\text{Hz m}$ is available [359]. According to the equation in [353] this yields the resonance frequency

$$f_{\text{res}} = \frac{N_c}{d_m} = \frac{1095\,\text{Hz m}}{68\,\text{mm}} = 16\,102.9\,\text{Hz}. \tag{Q.1}$$

The admittance magnitude peak of the FEM simulation lies only 10 Hz above that, at $f_m = 16\,112.7\,\text{Hz}$, corresponding to a 0.06 % difference. The offset between the FEM peak and the measured one at 16 169.3 Hz is 0.35 %. It means that concerning the resonance peak, all three values are in very good agreement, given that for frequency constants of piezoelectric materials a tolerance of up to 2-3 % is realistic. (With the preliminary PZT-8 material data (FEM: $f_m = 15.889\,\text{kHz}$) and the old frequency constant (eq. Q.1: $f_m = 15.735\,\text{kHz}$) a 3 % range is in fact covered by the offsets.)

What is not so well reproduced by the FE model is the position of the antiresonance and with it the width of the frequency interval between resonance and antiresonance. That interval allows to conclude on coupling coefficients which can

[5] Alternatively, this plane could have been explicitly declared a symmetry plane abolishing the need for any translational BCs and cutting the model size in half, but as the mesh size at this point is no issue, it was preferred to avoid unnecessary setup differences when going from the simple to the larger FE models, of which not all have a symmetry plane orthogonal to the rotation axis.

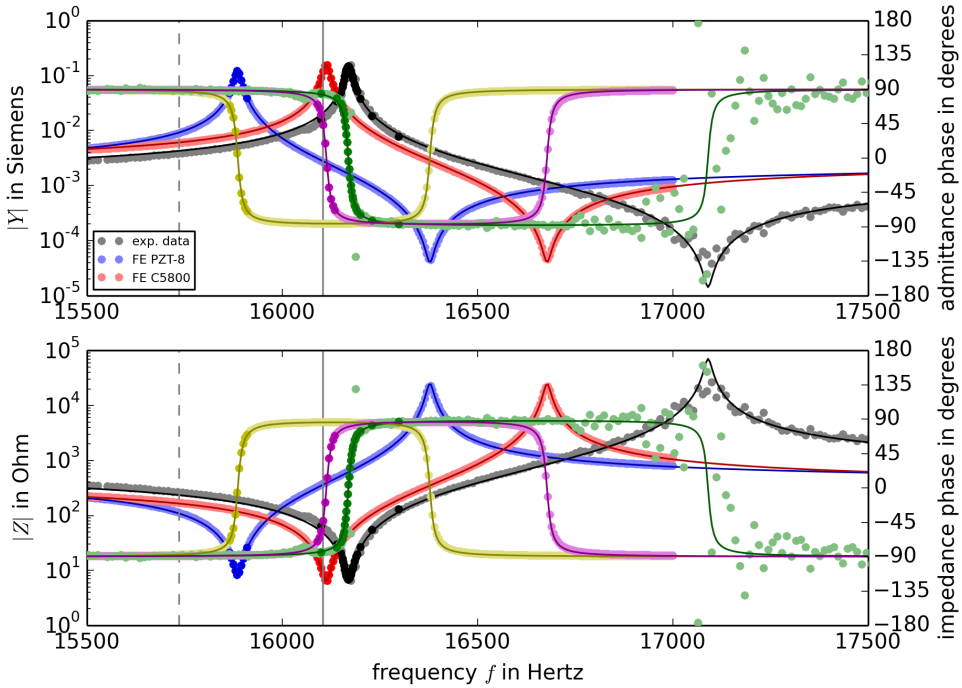


Figure Q.2 Frequency response of the simulated free transducer.
This plot compares the magnitudes and phases of admittance $|Y|$ and impedance $|Z|$ from the lab measurements (amplitude and phase in black & green, respectively; combined datasets 157 & 158) with two different sets of FEM simulation results, one based on the preliminary PZT-8 material constants (blue & yellow) and the other on the C5800 properties (red & magenta). The full colours mark the data subsets used for the Y-circle fitting. The dots in brightened colours show the complete data background of the corresponding frequency sweeps. The thin lines are the results of evaluating the BVD model (equation J.5) with the equivalent circuit quantities (R, L, C, C_0) gained during the Y- and Z-circle analyses of these datasets. As an additional comparison, resonance frequencies computed with the frequency constant N_c through equation Q.1 are indicated by the vertical grey lines; the dashed line corresponds to $f_{\text{res}} = 15\,735.3\,\text{Hz}$ and is based on $N_c = 1070\,\text{Hz m}$, while the one at $f_{\text{res}} = 16\,102.9\,\text{Hz}$ corresponds to $N_c = 1095\,\text{Hz m}$. Thus, the updated material properties and frequency constant bring the resonance frequencies computed by the FE model and equation Q.1 within a range of 10 Hz from each other, and as close as 70 Hz to the resonance peak observed in the laboratory.

be defined [353] as

$$k^2 = \frac{\text{electrical energy stored}}{\text{input mechanical energy}} = \frac{\text{mechanical energy stored}}{\text{input electrical energy}}, \qquad \text{(Q.2)}$$

and which are a measure of how well the material is able to transduce energy between the mechanical and electrical form. The Channeltech catalog [353] contains plots allowing to infer k going out from $(f_n - f_m)/f_m$. For the contraction and expansion mode of the hollow cylinder the relevant coupling coefficient is k_{31}. The C5800 material data set predicts (analytically) a coupling of $k_{31} = 0.32$ [359], and, according to Channeltech, the coupling in real transducers is likely to be a little bit higher, so values up to 0.34 or even 0.35 are possible [359]. In the case of the lab measurements, where the frequencies are $f_m = 16.112\,\text{kHz}$ and $f_n \approx 17.090\,\text{kHz}$, a value for k_{31} of circa 0.36 can be inferred. But as $f_n = 16\,678.51\,\text{kHz}$ in the FEM simulation, it

leads to the very low value of $k_{31} \approx 0.28$.

One more comparison can be made by plotting the admittance circles from measurement and FEM simulation into the same diagram, as shown in figure Q.3.[6] This plot shows multiple FEM simulation datasets corresponding to different mechanical damping ratios. The dielectric loss tangent had been kept fixed at the literature value [353] of 0.004. While the literature value for the mechanical damping ratio is 0.001 [516] (or $\tan\delta = 1/Q = 9.09 \times 10^{-4}$ with $Q = 1100$ [353]), the dash-dotted lines in the diagram show that values between 5×10^{-4} and 6×10^{-4} can best reproduce the observations on the real device. In particular, a damping ratio of 5.356×10^{-4} leads to a matching quality factor $Q = \frac{f_s}{f_{nB} - f_{mB}} = 920$, whereas a value of 5.764×10^{-4} will achieve an admittance circle with matching diameter, i. e. the right $G_{\max}$.

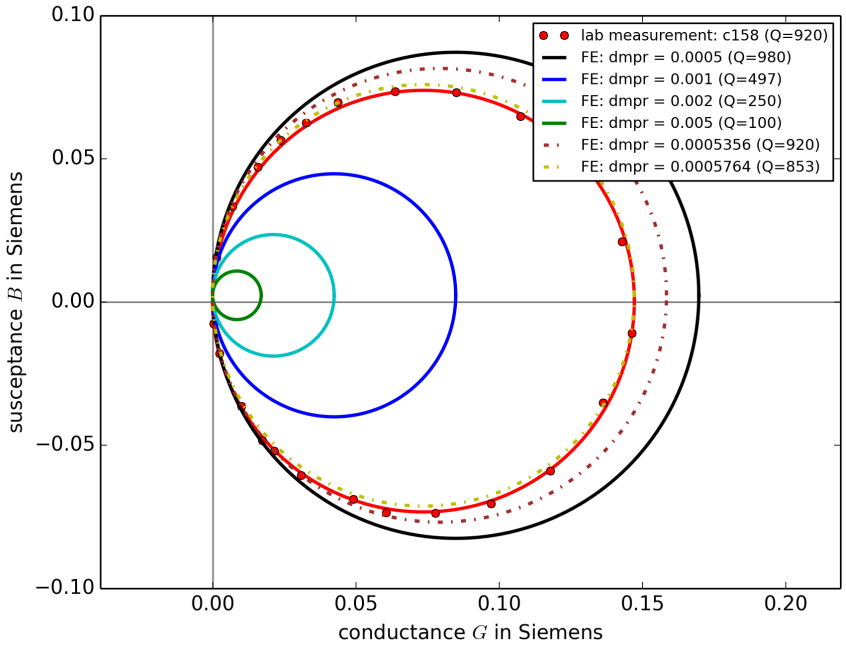


Figure Q.3 Admittance circles of the simulated free transducer.
Y-circles from the FEM simulation gained with various damping ratio settings (`mp,dmpr` command in Ansys) are plotted over the experimental data of case 158 (in red dots, the red line is the fitted BVD model). It can be seen that the Q-factor is directly related to the size of the admittance circle. The two dash-dotted circles represent two possible answers to the question which damping ratio in the simulation best fits the measured data. The yellow circle is the result of setting the damping ratio to 5.356×10^{-4} which leads to the exact same Q-factor of 920 if the formula $Q = \frac{f_s}{f_{nB} - f_{mB}}$ is used. The yellow circle represents a damping ratio of 5.764×10^{-4} which is the setting achieving the same circle size, i. e. the same $G_{\max}$.

In principle, the dielectric loss tangent could be considered as another degree of freedom in the Ansys FE model allowing for tuning in order to more closely match the characterisation data from the real transducers, or to discuss the relation of literature and observed values at different power levels. The dielectric losses affect the size ratio between the Y- and the Z-circle. This is shown in appendix J.5 where the admittance

[6] An analogue plot for the impedance circles is not possible due to the noisy current probe data around the antiresonance, where the current amplitude goes through a minimum.

and impedance responses of the FE model and the two equivalent circuits, the BVD and the complex circuit model, are examined more closely. In the current case this is not possible because the electric characterisation data recorded in the lab has a degraded signal-to-noise ratio around the antiresonance. But in the future it should be possible[7] to characterise the transducers in terms of both mechanical and dielectric damping before glueing them together with the other resonator assembly parts, and thus making available a better material data basis for improving the FE model of the whole resonator.

Q.2.3 Comparison: water-filled resonator in Atila and Ansys

The first complete FEM simulation of a liquid-filled resonator geometry discussed here is a reproduction in Ansys of the FEM simulation conducted by Cancelos [69] with the software Atila®. It serves the purpose of a code-to-code validation. The mesh is shown in figure Q.4 along with a comparison of the sound pressure response. The good agreement of the two independently composed FE models is a strong hint that the two models have been set up as intended because any mistake would have to be a common one.

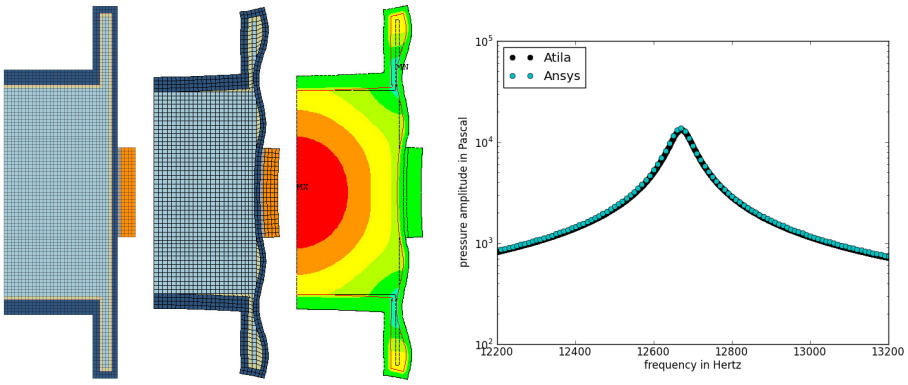


Figure Q.4 FE model of Cancelos' glass resonator.
The plot on the left depicts the FE mesh implemented in Ansys, and it has to be compared to figure 16 on page 42 in [69]. The maximum element size setting is 1.5 mm leading to a mesh of about 3000 finite elements. The boundary conditions are the same as as with the free transducer, i. e. zero displacement along z for the nodes in the midplane, grounding of the inner electrode, and 1 V applied to the outside electrode. The two pictures in the middle show the deformation and the pressure mode shape at the resonance frequency of 12 680 Hz. The deformations are scaled with a factor of 10^4. The plot on the right hand side shows the direct comparison of the sound pressure response picked up at the central node of the FE mesh with the corresponding dataset furnished by Cancelos. Both frequency and amplitude of the resonance peak are in almost perfect agreement.

Q.2.4 Simulating resonator Nº 5

Resonator Nº 5 (in the RPI series of resonator exemplars listed in appendix I.4) is the one that was chosen for thorough examination because of being manufactured

[7] The low signal-to-noise ratio is due to the minimum in the current amplitude at the antiresonance. But had the measurement been taken with the transformer in place between amplifier and transducer, then the current amplitude minimum would not have been that deep, as can be concluded from a look at figures O.20 and O.21.

most symmetrically and cleanly, and because its geometry matches Taleyarkhan's sketch (fig. D.1) more closely than the other RPI exemplars. It was assumed to furnish the best available validation scenario, to present an occasion where a well-matching FEM geometry can be most easily created and where a close similarity of sound pressure and glass wall displacement maps can be most probably achieved. However, the match between FEM simulations and lab measurements turned out to be rather imperfect. In the following, this FE model and the matching and not matching features of the compared data fields will be discussed.

The FE mesh

The model of resonator 5 (shown in figure Q.5) has several new features with respect to the FE model discussed before (figure Q.4). The materials steel and aluminium are introduced for modelling the hydrophone and the flange at the glass wall's top rim. Silicone is used for connecting the three structural parts of glass and aluminium. The artificial fixation material forms a bearing for the aluminium flange with zero displacement BCs at the lower surface of the bearing. A last additional feature is the free surface of the liquid. Here, a zero amplitude BC is applied to the nodal sound pressure DOF. With the parameter for the longest allowable element edge set to 1 mm the mesh has around 6×10^3 elements and 10^4 nodes in the freely meshed which had been used by default. Alternative versions with mapped mesh creation routines (see fig. Q.12) yielded circa 8.5×10^3 elements and 12×10^3 nodes.

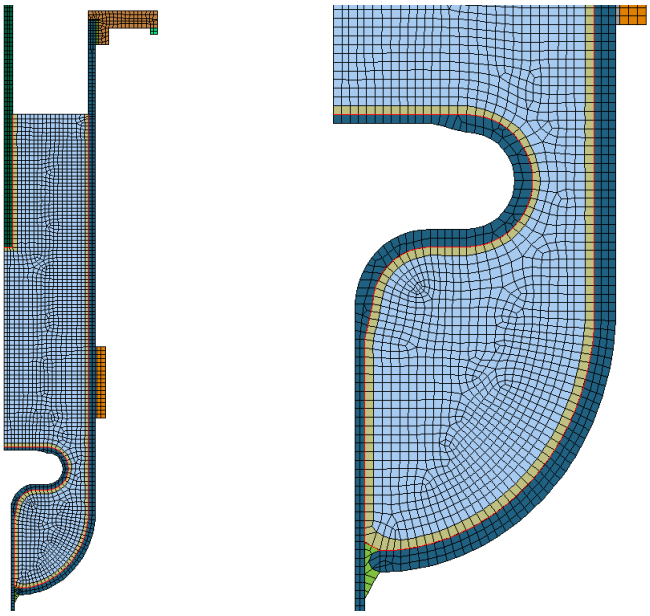


Figure Q.5 FE mesh of resonator № 5
This FE mesh of the opened resonator 5 served the purpose of validating the FE model. The hydrophone is modelled as a massive steel rod. The fixation of the mesh occurs through a bearing consisting of a ring of soft material glued under the aluminium flange near its outer rim. For the plot on the left the element size parameter has been raised (from the otherwise used setting of 1 mm) to 1.5 mm to allow better visibility.

Simplifications and differences of the FE model with respect to the real-world counterpart

The most important differences between the FE geometry and the real resonator are: the 2D-axis-symmetric model assumes perfect rotational symmetry whereas in reality there are small deviations, in particular concerning glass part shapes and thicknesses of epoxy and silicone glue layers. Next, the model geometry is laid out as a composition of added and subtracted rectangles, circle segments, trapezoids etc., and some linearily stretched versions of these base forms. These forms try to approximate a glass shape, that in some corners (e. g. the piston curvatures and inside dimensions) cannot be readily determined with normal machine shop tools (X-rays would be a solution). Besides the unintended shape irregularities there are additional features breaking the rotational symmetry: four boreholes and bolts in the aluminium flange, the pill microphones on the glass a little distance above the transducer, and the solder joints on the outside and inside surfaces of the transducer, of which the latter one protrudes into the 0.5 mm thick epoxy coupling layer, so that it can be assumed to introduce some inhomogeneity into the force transfer between transducer and glass. Another difference lies in the hydrophone fixation, not modelled in the FE mesh, but consisting in the special aluminium top head lying on the flange counterpart with the rubber O-ring in between. The four connecting bolts were present and weakly tightened for the purpose of degassing (see appendix O.4). Finally, the real chamber was suspended by four wires attached to the lower ends of the four bolts.

Comparison of the pressure field

Figure Q.6 shows a comparison of the frequency response picked up with the hydrophone inside resonator № 5 with several FEM simulations differing only in their damping settings. Generally, the FE model is able to reproduce the three resonances recorded on the real device in the frequency interval of interest, although substantially shifted. Secondly, it exhibits a realistic peak Q-factor for the second resonance (the SF working point), not only when a global mechanical damping ratio of 1.43×10^{-3} (deduced from the experimentally observed $Q = 700$) is applied, but also when literature damping ratios are attributed to the individual materials. On the one hand, it can be seen that some degree of uncertainty on the exact quantity of the transducer losses, which are comparatively low with respect to the other employed materials, does not pose a problem because the impact of a variation on the overall behaviour of the FE model is small. On the other hand, neglecting the material-dependency of damping at all, does in fact change the peak landscape.

The sound pressure map, where the hydrophone frequency responses at many probing positions along the central axis are compiled to form a contour plot, is more telling in the context of validation than just one single response because the mode shapes of the standing pressure waves in the liquid are revealed. Based on the insights described above and in figure Q.6, the setting with individual literature-based material damping ratios and $\tan\delta = 5.356 \times 10^{-4}$ for the piezoceramic was chosen for presentation in figure Q.7. Comparing the two maps, one can say that the FE model is able to capture the basic features of the resonator and even some

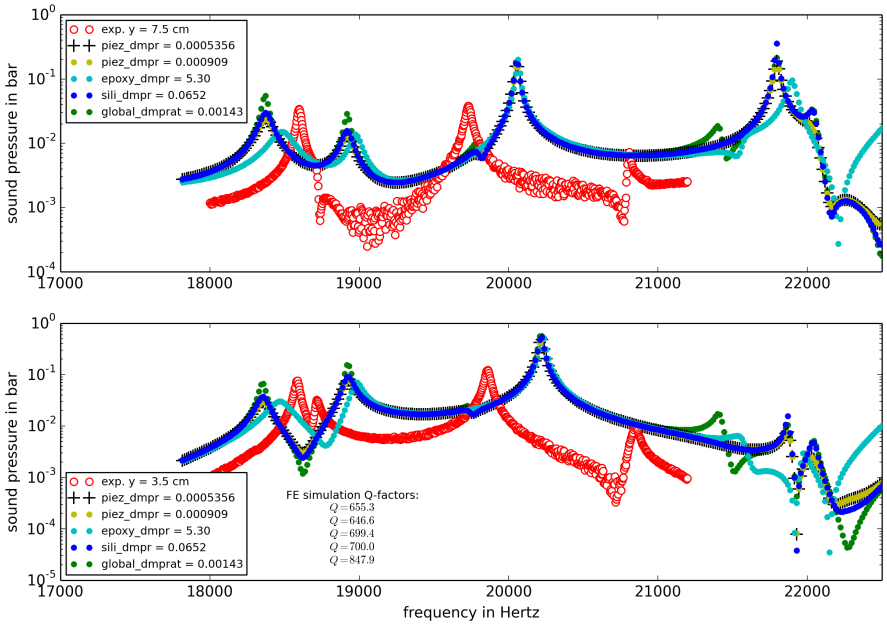


Figure Q.6 Hydrophone response and different damping settings
These plots show the hydrophone response in a comparison of lab with FEM data for two different vertical positions of the probe. The red circles in the background represent the empirical measurement data taken at $y = 7.5\,\mathrm{cm}$ (top diagram) and $y = 3.5\,\mathrm{cm}$ (bottom). The FEM simulations of each case were repeated several times with different damping settings. In the base setting the literature values for the damping ratio in each material were employed, i. e. 5×10^{-4} for glass, 1×10^{-4} for aluminium, 0.03 for epoxy, 0.075 for silicone, and $\tan\delta = 9.09 \times 10^{-4}$ as well as $\tan\delta_{\mathrm{diel}} = 4 \times 10^{-3}$ for the mechanical and dielectric loss tangents of the transducer material. The mean pressure signal loaded onto the nodes forming the hydrophone tip surface under these conditions is plotted in yellow. The black crosses show what happens when $\tan\delta$ is changed to 5.356×10^{-4} for the C5800 ceramic. It is the value which had been determined as the one able to reproduce the experimentally observed Q of the free transducer. The green curve is the result of a different approach, where one single mechanical damping ratio of $\tan\delta = 1/700 = 1.43 \times 10^{-3}$ is applied uniformly across the whole FE model (in conjunction with $\tan\delta_{\mathrm{diel}} = 4 \times 10^{-3}$ for C5800 as for all other cases). The Q-factors written into the diagram were inferred from analysing finer-resolution close-ups of the second resonance (near 20 kHz), and they seem to indicate that all these three simulation cases match the experimentally observed Q-factor fairly well. The remaining two cases are the results of forcing the match by 1D parameter tuning procedures. For the cyan curve that particular damping coefficient for epoxy was searched which would yield a Q of 700 while damping in the other soft material, silicone, was switched off. For the blue curve it was the other way round. This may be taken as the upper limits of damping supported by the experimental observation for the two materials where literature values have the weakest justification. Looking at the other features apart from the central resonance, in particular the cyan and green curves reveal that not only the relative sharpness of the peaks is affected, but that there are peaks appearing in one case and completely lacking in the other. Finally, one can conclude that the question of the exact transducer losses is a detail without heavy impact, but that the question whether individual material damping ratios should be accounted for or not, on the other hand, is more important.

of the details: there are three resonances, the vertical mode shape of each one is roughly reproduced, and the resonance frequencies agree with offsets up to 1 kHz. An interesting feature is that the first resonance seems to consist of two interacting modes which results in a split into two peaks in the lower two maps. In the upper map the second mode is just barely visible as a side peak or shoulder in the upper part of the image. As the first two rows of images represent data gained on the very same resonator – a time period of four months lies between the two recordings but there were no intended changes in setup – this is an indicator of the sensitivity of that device. The FEM simulation exaggerates the split, the twin peaks are further apart, the gap in between is not as sharp as exhibited by the second experimental set. But the mode shapes of the twins, that the left one has a node at $y = 6\,\mathrm{cm}$ and that the right one by a thin tail seems to extend all the way up to the free liquid surface at $y = 8\,\mathrm{cm}$, these properties are well reproduced by the simulation. Other features not well captured by the FEM simulation are the asymmetry of the slopes along the frequency axis of the twin peaks and the third resonance, the frequency offsets, the lack of the small and inclined ridgeline existing in the upper map between -2 and $0\,\mathrm{cm}$, and most importantly, that the relative peak heights exhibited by the three resonances are not correctly mirrored. But taking into account the lessons learnt through the sensitivity study following in this appendix chapter, it could actually be considered unrealistic to expect a perfect match.

A last remark has to be made on the sound pressure developed per unit of driving voltage: in the measurement row furnishing the uppermost map in figure Q.7 a pressure of $1 \times 10^5\,\mathrm{Pa}$ was achieved with a driving amplitude of 1.8 V, that is 0.54 bar per Volt. In the second plotted set it is $\approx 0.1\,\frac{\mathrm{bar}}{\mathrm{V}}$ (0.24 bar with 2.3 V). The FEM simulation shows around 25 bar created with a driving voltage of 100 V, i. e. $0.25\,\frac{\mathrm{bar}}{\mathrm{V}}$, so it is roughly within the same range. A problematic aspect is that these highest pressure gains were achieved in different resonances. If looking exclusively at the second resonance, which is the SF working point described by Saglime [390], in the three datasets of figure Q.7, then values of $\approx 0.1\,\frac{\mathrm{bar}}{\mathrm{V}}$ and $\approx 0.05\,\frac{\mathrm{bar}}{\mathrm{V}}$ have to be compared[8] with the afore-mentioned gain of $\approx 0.25\,\frac{\mathrm{bar}}{\mathrm{V}}$. That the peak heights differ by less than one order of magnitude can almost be considered to be a surprisingly tight match, taking into account on the one hand the performance sensitivity (discussed below) and on the other hand how the peak heights are influenced by various material damping ratios.

Fitting losses by Q-factor matching

If one had several trustworthy material models, including a precise quantification of the damping behaviour applicable to the examined geometry, and only a single material with an unreliable damping coefficient, then one would be able to use the experimentally gained knowledge of the Q-factor of a resonance to infer the strength of the unknown dissipation mechanism. One could ask: which amount of damping in the questionable material is necessary to explain the observed Q? This was done

[8] The pressure gains are best inferred from the pressure profile plots. For the three datasets displayed in figure Q.7, the vertical profiles are given in figures O.24, P.22, and S.1 on pages 387, 414, and 467.

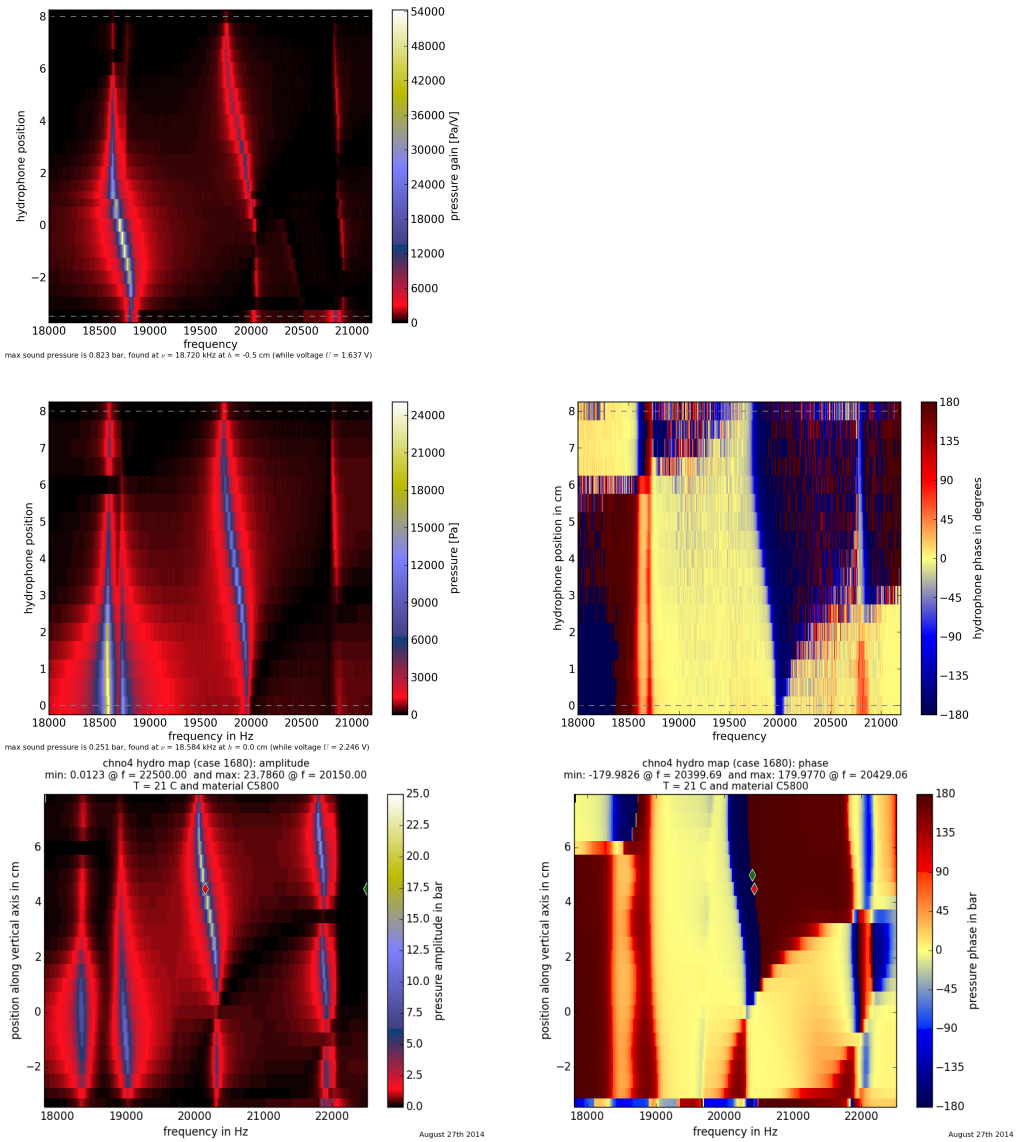


Figure Q.7 Sound pressure map of resonator № 5

These colour maps show the comparison between laboratory (first two rows) and simulation (lower row) data for the sound pressure measured at the hydrophone tip. Amplitude plots are on the left, phase signals on the right. In both the lab and the numerical experiment the maps have to be created by assembling many frequency sweeps generated in series after vertical shifts of the hydrophone position. The grid of measurement locations has a steps size of 5 mm and is the same in both cases. In the first dataset, the highest position coincides with the liquid surface and the lowest position is 0.5 mm away from the piston. In the simulation the extreme positions are 2 mm further inwards on each side. The experimental data in the second row does not span the whole vertical area, but contains the phase information which has not been recorded earlier. The measured hydrophone phase data have been cycled downwards through the $[-180°, 180°]$ interval by subtracting 170° in order to achieve matching colours with respect to the FE data plotted below and to allow a proper comparison of structures in the colour landscapes. The comparison reveals that mode shapes can be captured quite well by the FE simulation, but not their relative heights on the pressure scale nor the exact positions on the frequency axis. Figure S.2 in appendix S can be added to this comparison. It represents a modified FE simulation where the alternative literature value for the Young's modulus of the epoxy (see table Q.3) has been tried out. The map changes indeed, however, with respect to the similarity between simulation and reality the situation is not changed.

for epoxy and silicone, the two soft polymer materials. It is possible, because the FE model exhibits a lower damping and a higher Q (the value is about 2500) than the observed $Q = 700$ of the second resonance (see figure O.27) if the losses of epoxy and silicone are switched off. The blue and cyan curves plotted in figure Q.6 show the results of tuning the damping ratio in either epoxy or silicone alone up to the point where the second resonance has the right Q. So, it is a damping ratio of 5.3 in epoxy or 6.5×10^{-2} in silicone that could explain the observed Q in the absence of dissipation in the other polymer.

An attempt was also made to use the Nelder-Mead algorithm to fit[9] the Q-factors of both the second and third resonance (the latter has $Q = 957$) at the same time. The two degrees of freedom in this case were one damping ratio for all hard structural materials (except the piezoceramic where $\tan\delta = 5.356 \times 10^{-4}$ has been taken) and another one for the two soft polymers. But no solution could be found. In this minimisation procedure, the Q of the third resonance would not rise above the Q of the second resonance. The best point found by the algorithm was $Q_2 = 813$ and $Q_3 = 714$ with $\tan\delta_{\text{hard}} = 1.5 \times 10^{-3}$ in epoxy and $\tan\delta_{\text{soft}} = 3 \times 10^{-9}$ in silicone. The first resonance was never considered for such trials due to it being formed by the interaction of two peaks. In the FEM simulation it does not matter at which point along the central axis the Q-factor of the sound pressure peak is determined, it is ever the same value.

The examined devices all exhibit many resonances with very different mode shapes. Some resonances are more determined by the fluid, some are dominated by the structure, e. g. internal modes of the glass wall. Different modes induce different amounts of stress in different locations, so it can be assumed that e. g. damping in silicone joints does not affect all resonances in the same way. Indeed, when varying damping ratios of single materials in the FE model, it can be observed that the different peaks are not broadened equally. In principle, the complex structures in the sound pressure and glass wall displacement maps hold a reservoir of many observables, the Q-factors are just one sort, allowing their use as degrees of freedom in minimisation procedures with the purpose of fitting material or also geometry parameters. However, this method of using the resonance Q-factors for determining damping coefficients was not further pursued in this place, also in the light of the too many unknown deviations between model and measurements, and because of the design sensitivity issues discussed below. The FEM sound pressure map needs to look different, much more similar to the real-world counterpart first, the resonances should be pushed into their right places first, before further Q-factor fitting makes real sense. The sound pressure amplitude and phase maps are feature-rich and offer themselves for model calibration tasks, but in order to achieve a realistic outlook, a model improvement agenda should be pursued in a stepwise approach beginning with smaller and simpler systems before calibrating entire resonator models.

[9]The goal of the downhill-simplex optimisation run was to minimise the sum of (normalised) square offsets of the two Q-factors, i. e. to minimise $\sum_{i=2}^{3}[(Q_{i,\text{FE}} - Q_{i,\text{exp}})/Q_{i,\text{exp}}]^2$.

Comparison of the radial displacement field

Since the glass surface of the resonator was scanned with the displacement pickup needle, a second 2D colour map can be shown for comparison, a map of the radial displacement of points on the surface. The comparison is made in figure Q.8 where it can be seen that this map is richer in structure because the cylindrical glass wall is excited to higher-order modes. Just as in the pressure maps, the relative amplitudes do not match, and there are of course the same frequency offsets. The split dividing the first resonance up into several interacting modes is visible, and again the spread is much larger in the FEM data. At least in both cases there is a structure with three main bellies along the vertical profile. The mismatching phase datasets on the right side in figure Q.8 show even more than the amplitude maps how far off the simulation is from the measurement at the moment. However, it can be remarked that making available the glass wall displacement data with its rich structure, adds a perfect set of calibration, benchmarking, and validation data to the electrical and hydrophone signals which had been available before.

It had been pointed out in figures O.23 and O.24 (pages 386 and 387) that the comparison of the sound pressure map with the data of Saglime ([390], reprinted in figure I.1, p. 284, recorded at 19.9 kHz) shows that the resonance at 20 kHz used to be the working point of former RPI SF trials. Therefore, this resonance deserves a closer look as presented in figure Q.9 which brings together the FEM data on deformation and pressure. Already the three resonator cuts in the upper left corner of that figure explain a lot. They show that the cylindrical main glass wall is divided up into two segments, the lower one extending from the height of the piston upwards to the upper rim of the piezo ring, that segment moves inwards and outwards in phase with the transducer, and another slightly larger segment above oscillating out of phase. The third cut with the colouring according to the pressure mode shape makes clear that regions of high pressure can be found on same heights with contracting glass walls and low-pressure areas are behind widened glass walls, i. e. the radial movements of liquid and surrounding glass wall at a given height are in phase, and along the radial coordinate only the fundamental acoustic mode is excited in the fluid. The front plate of the glass piston is not only moving up and down, it is also excited by a bending mode. But on average, the piston front, as another boundary of the main liquid volume, oscillates inwards and outwards in phase with the transducer.

The colour map in the upper right corner of figure Q.9 shows again the displacement amplitude of the glass wall. But this time it is not only the section that has been scanned with the pickup needle. This map covers the mesh of the whole resonator side wall, from the top rim including a piece of the aluminium flange down to the silicone bead sealing the lower outlet and holding the piston in place. The liquid level, the transducer rims, and the piston position are indicated for orientation by horizontal lines and the grey shading. The plot shows the normal displacement, not the radial movement, of the FE mesh nodes on the inside surface of the resonator's wall, i. e. the oscillation orthogonal to the surface. This plot reveals that the two large segments of glass wall moving in phase with the liquid behind it (and the segments against each other) at 20.1 kHz seem to be the exception in that frequency

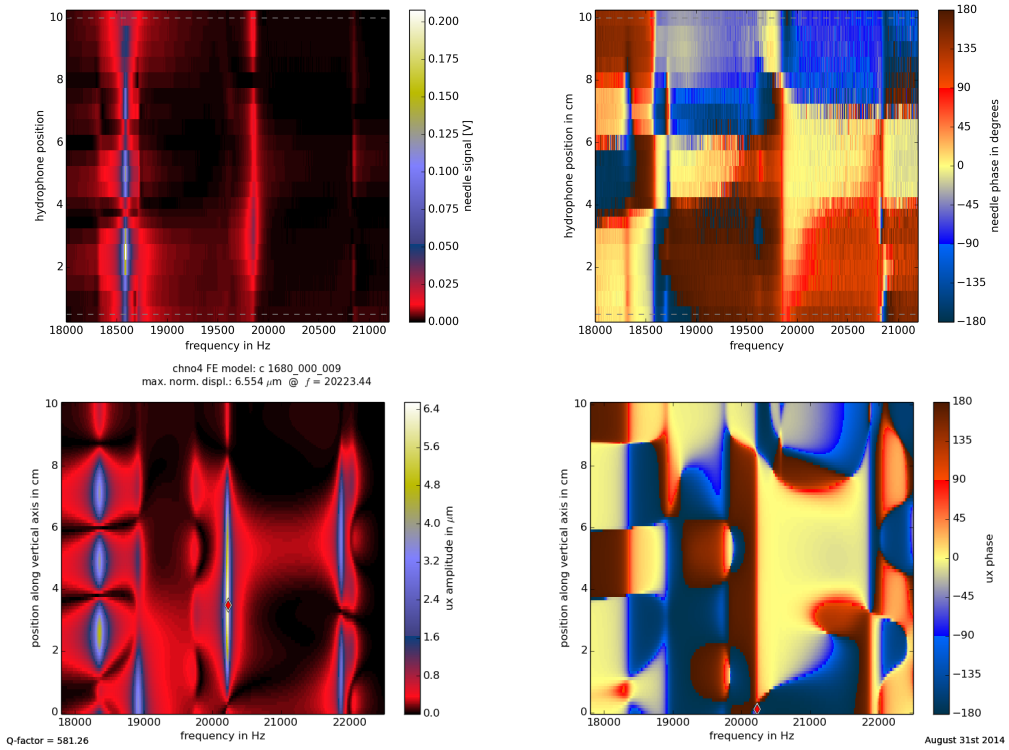


Figure Q.8 Radial displacement map of resonator Nº 5
This is a comparison of the radial wall displacement data between laboratory measurements with the pickup needle (top) and FEM simulations (bottom). The scanned area is the outside surface of the main glass cylinder between the transducer (its upper edge defines the origin $z = 0$ of the vertical axis) and the aluminium flange. The comparison reveals a worse situation as with the pressure maps because only two of the three mode shapes look similar. Looking at the amplitude plots on the left, the mode shape of the 3rd resonance consists of three antinodes in the pickup needle map but the counterpart in the FEM results displays a two-hump structure with a shorter-wavelength structure close by. For the other two resonances the situation is similar as with the pressure maps: the mode shapes are roughly reproduced, but not their relative amplitudes, nor details appearing in the valleys in between, nor exact resonance frequencies. Similar to the pressure data, the 1st resonance consists of several interacting resonances being relatively closely packed in the recorded lab data and further apart in the FEM data. In the needle measurement there is a strong three-and-a-half-belly resonance with two weak structures on either side, whereas in the FEM map a strong multi-belly resonance has a neighbouring resonance to the right which is weak in the upper part but has a comparably strong amplitude at the lower end. A positive remark may be made pointing out one detail feature wherein the two maps match: the structure to the left of the 2nd resonance is present in both cases, although with much weaker amplitude in the empirical data. The match of the phase data (right column) is quite poor. Regions of same colour in one map appear differently in the other one. For this presentation the phase of the pickup needle voltage signal had to be cycled upwards by 130° in order to match the red-to-white transition across the 2nd resonance displayed by the FEM data. This manipulation of the phase offset has to be considered in the context that the absolute phase transfer function of the hydrophone and its electronics is at this point unknown.

range, where the other resonances exhibit wave patterns of much shorter spatial wavelengths with nodes of zero displacement not further apart than 3 cm. It can be taken as a sign that the fluid forces its motion on the glass at that resonance more than at other frequencies. But the fact that none of the standing wave patterns in the glass seems to ignore the acetone level at 8.1 cm tells us that probably none of the wave patterns can be described as a “pure glass mode”. To label some resonances

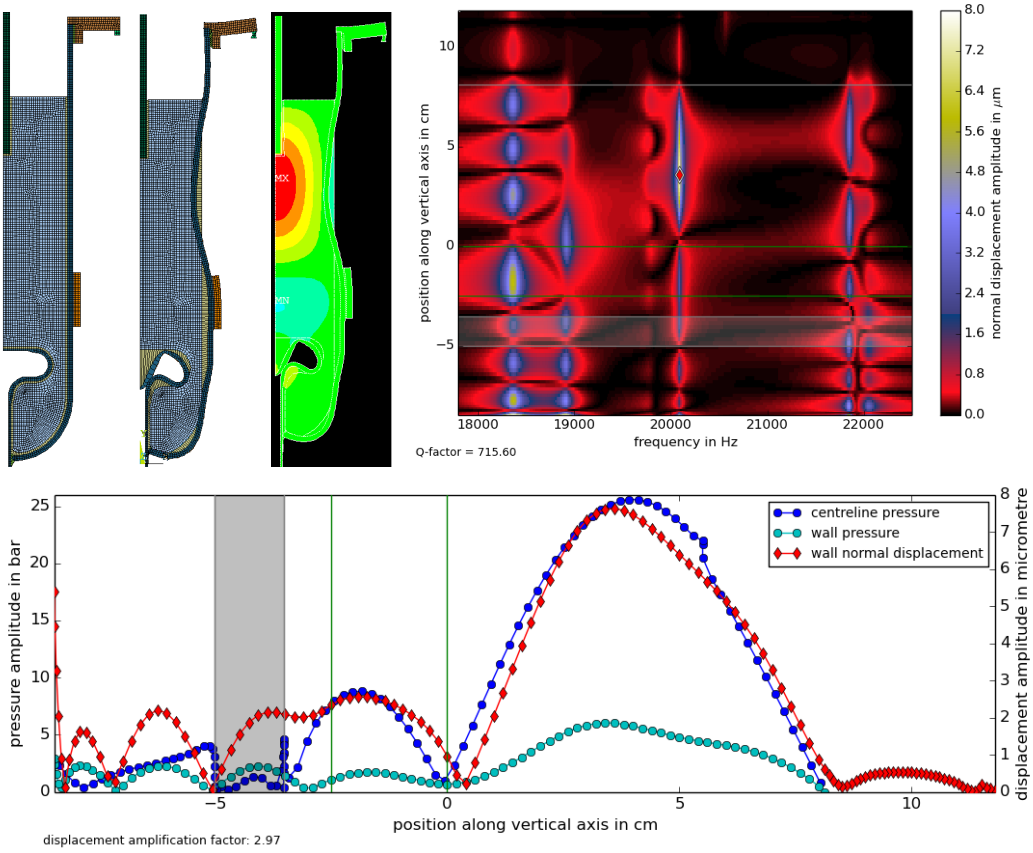


Figure Q.9 Pressure and deformed shape at the second resonance
In these pictures the pressure field and the structural deformation are brought together for a close look at the 2nd resonance at 20.01 kHz. By its mode shape this resonance can be identified as the working point used for past SF experiments. The three pictures in the upper left corner show the undeformed mesh, the deformed mesh, and the pressure field across the mesh. It can be seen that the pressure amplitude on the piston surfaces and along all other boundaries of the fluid domain is quite low, nowhere is it higher than 20 % of the pressure in the centre. In the deformed mesh plot the displacements are scaled up by a factor of 800. Here and in the radial displacement amplitude colour map in the upper right corner it becomes clearly visible that the glass wall all across the long range from 2 to 6 cm moves inwards and outwards much further than the piezo transducer does. The transducer's extension is marked by the green horizontal lines in the colour map. The exact amplification factor between the motion amplitude of the transducer and the maximum seen by the glass 3.5 cm above is 2.96. The plot at the bottom collects the mode shapes of pressure and displacement in one diagram. The pressure along the resonator's central axis and further down the inside boundary of the fluid domain where it coincides with the piston's surface is shown in blue. The pressure along the outside of the fluid domain boundary, i. e. the inside of the glass wall, is shown in cyan. The red diamonds represent the glass wall's normal displacement. The shaded area indicates the position of the lower piston, and the green lines show the extension of the transducer. The hydrophone position at 5.5 cm can also be inferred from the edge in the pressure profile. All plots in this figure stem from the same simulation with this hydrophone position; it is the sixth in the series of simulations used to compile the pressure map in figure Q.7.

as “acoustic modes of the liquid volume”, where the glass motion follows the liquid and the latter oscillates in one of its eigenmodes, and others as “modes of vibrating structure”, where the liquid follows the glass and the latter oscillates in one of its eigenmodes, e. g. to label the second resonance as a “resonance dominated by a liquid eigenmode” may sometimes be helpful, but all too often it could lead to an over-simplification. It is better to keep in mind that the complete assembled resonator has other eigenmodes than its separate parts would have. Think of what happens when one touches a guitar string with the finger right in the middle. This affects only every second eigenmode, the ones with an antinode in the middle. Releasing the finger could be seen as putting the two parts together again: the whole string can exhibit all the eigenmodes of the two separate halves plus additional new modes. Does it mean that the eigenmodes of the liquid-filled resonator are the eigenmodes of the liquid combined with the modes of the structure plus additional modes? No, this would be wrong. One should not overlook that there is a difference between just touching the guitar string in the middle or clamping it down really hard and completely decoupling the halves. Only the hard clamping allows the two halves to oscillate independently in different modes at any phase. In the case of the softly touching finger the second half of the string still functions as a boundary condition of sinusoidally varying force to the first half. In our resonator, the liquid and the structure have each other as boundary conditions, they can locally slam against each other which creates a pressure antinode at the wall, or they can move in line so there is a pressure node on the structure surface. The cyan line in the plot at the bottom of figure Q.9 shows clearly what the blueish shade near the glass wall in the upper half of the resonator in the pressure mode shape plot in the same figure is only hinting at: that the sound pressure amplitude near the glass wall is nonzero in most places but still substantially lower than the amplitude in the antinode in the centre of the chamber. A resonator for SF experiments could in a certain way be seen as ideal if the structure surfaces where coincident with sound pressure nodes anywhere, because this would completely exclude cavitation on the walls at any driving amplitude. There would be no force transmission between the fluid and the solid, and both parts would in principle be able to exhibit the same oscillation mode shape in the absence of the other part. But in reality some level of coupling is desired for the transducer to be able to excite the acoustic mode in the fluid via the structure, and it is no problem as long as the cavitation threshold is not breached anywhere on the interface. In short: it is better to keep in mind that all oscillation modes of the resonator are a property of the whole assembled ensemble of structural parts and the fluid domain.

After this look at the measured and simulated pressure and displacement frequency responses, it has to be admitted that the exact resonance frequencies and relative peak amplitudes do not match so far. There are substantial differences in detail visible in the structure-rich colour maps showing amplitude and phase over position and frequency for the centreline pressure and the wall displacement. Only the general sequence and mode shapes of the three resonances are mostly in agreement.

Already at this point it can be said that a design approach starting with dimensioning a cylindrical liquid volume so it alone would exhibit a desired mode shape, and in the second step constructing a bounding box of glass or steel around it,

will not be the approach leading to the best results because the fluid and structure domains need to interact, and the walls cannot be built infinitesimally thin.

Comparison of admittance and impedance

The electrical side of the transducer offers additional observables for comparing the FEM simulation with its real-world counterpart. In the lab, the harmonic current and voltage signals feeding the transducer electrodes can be tracked, and the frequency-dependent admittance $Y(\omega)$ and impedance $Z(\omega)$ of the resonator can be deduced from these signals. In the harmonic FEM calculation, the voltage is the given vibration-exciting boundary condition and the current flowing in and out of the transducer electrodes can be deduced from the reaction force "electrical charge" building up in the corresponding nodes of the FE mesh. Technically, the total charge Q of an electrode can be gained by summing over nodal charges q_i, or alternatively, Ansys offers the reading of the total charge via *master nodes* [7] if a coupled set of nodes has been defined before. In the case of a stationary harmonic oscillation the relation between charge Q and current I is given by

$$I = \dot{Q} = i\omega Q \qquad \text{whereby } Q, I \in \mathbb{C}. \tag{Q.3}$$

This way, the admittance and impedance frequency response of the FE model can be gained. Again, the response functions allow the fitting of equivalent circuit models. A corresponding admittance comparison for resonator 5 and its FE model is plotted as magnitude and phase diagrams in figure Q.10. Plots of the same datasets in the complex Y- and Z-planes are displayed in figure Q.11. The results of the analyses of the Y- and Z-circles are given in table Q.5.

Looking at the frequency response of the admittance in figure Q.10, there do not seem to be many aspects in which the experimental data and the simulation agree. Of course there are the already discussed frequency shifts. In terms of shapes of the curves, only the second resonance looks similar. In the case of the third resonance, the depth of the dip in phase is at least comparable. In the case of the first resonance, there is not only the difference in how far apart the two halves of the structure are, it seems also that they are of very different degrees of sharpness and following completely different patterns in the phase function.

Taking the plots of admittance and impedance circles into consideration can put the different types of frequency sweep curves into relation. While the FEM simulation seems pretty far apart from the experimental data in the Y-circle plot of figure Q.11, it is a glance back at figure O.22 which shows that the diameter of slightly less than 10 mS of the simulation-based Y-circle is not so far outside the range of the lab recordings going from slightly more than 10 to above 50 mS. The maximum conductivity, the equivalent circuit quantity R, the degree of damping and the mechanical Q are all tightly related. For the first resonance, the lab observations for Q range from 270 to 1000 and for R from 74 to 18 (see table O.4). Here, the FEM dataset properties with the below-range Y-circle diameter translate directly into an above-range resistivity $R = 1/G_{\max} = 125\,\Omega$ and a below-range Q of ≈ 200. For the 2nd and 3rd resonance the FEM data with Q-factors of 650 and 500 compares to 500 (tab. O.5) and 640 (fig. P.17), so they are in a similar range with each other,

Table Q.5 BVD quantities for the FE model resonances of resonator 5

For each of the four resonances exhibited by the FE model of the opened resonator 5, the admittance and impedance circles can be plotted and analysed allowing the deduction of a BVD circuit model which would reproduce the same pattern. This was done for the FEM simulation with the hydrophone at 3.5 cm. The Y- and Z-circles are plotted in appendix S (figures S.3 & S.4). This table gives the characteristic frequencies and the BVD circuit quantities. Since the first resonance is split into two, labelled "1a" and "1b", there are four resonances of which the BVD analysis results are listed in the four last columns. However, considering the current phase curves plotted in figure P.23, the first resonance can also be interpreted as a wide structure with an overlaid narrow artefact. Therefore, in an additional column, labelled "1", the analysis data is given taking the Y-circle of the resonance from "1a" and the Z-circle of the antiresonance from "1b". The change affects C_0, k, and M.

quantity	unit	formula	res. 1	res. 1a	res. 1b	res. 2	res. 3
T	°C		21	21	21	21	21
f_{mB}	Hz		18313.3	18313.3	18872.9	20202.1	21846.9
f_m	Hz		18343.4	18343.4	18894.2	20209.9	21855.3
f_s	Hz		18359.5	18359.5	18919.1	20218.3	21870.4
f_r	Hz		18382.1	18382.1	–	–	–
f_{nB}	Hz		18408.6	18408.6	18957.1	20233.2	21890.3
f_a	Hz		–	18463.5	–	–	–
f_p	Hz		18961.9	18489.1	18961.9	20238.4	21887.3
$G_{\max}$	mS		8.03	8.03	2.42	3.59	2.67
B_s	mS		3.08	3.08	1.94	2.58	3.02
Q_m		$\frac{f_s}{f_{nB}-f_{mB}}$	192.6	192.6	224.7	652	504
Q_e		$\frac{B_s}{G_{\max}}$	0.384	0.384	0.803	0.718	1.132
R	Ω	$\frac{1}{G_{\max}}$	124.5	124.5	412.9	278.9	374.2
L	mH	$\frac{Q_m R}{\omega_s}$	207.8	207.8	780.4	1430	1373
C	nF	$\frac{1}{Q_m R \omega_s}$	0.362	0.362	0.091	0.043	0.039
C_0	nF	$\frac{f_r^2}{f_a^2-f_r^2}C$	–	40.78	–	–	–
C_0	nF	$\frac{f_s^2}{f_p^2-f_s^2}C$	5.42	25.52	20.01	21.83	24.95
C_0	nF	$\approx \frac{B_s}{\omega_s}$	26.73	26.73	16.35	20.27	22.01
k		$\sqrt{\frac{f_p^2-f_s^2}{f_p^2}}$	0.250	0.118	0.067	0.045	0.039
k		$\sqrt{\frac{f_a^2-f_r^2}{f_a^2}}$	–	0.116	–	–	–
M		$\frac{k^2 Q_m}{1-k^2}$	12.84	2.73	1.018	1.293	0.779
Γ	nF m^{-1}	$\frac{Cd}{A}$	0.203	0.203	0.051	0.024	0.022

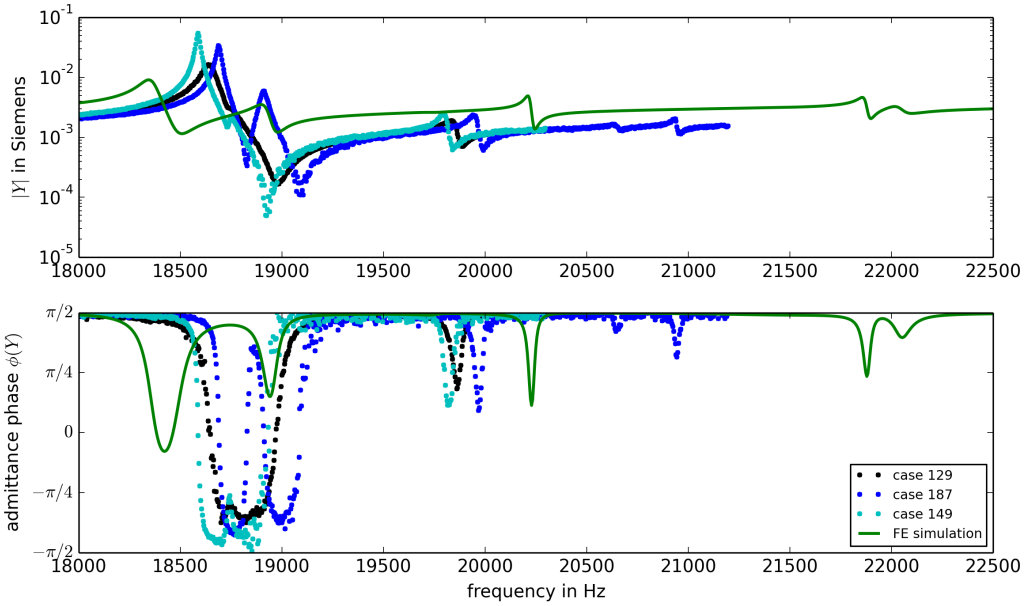


Figure Q.10 $Y(f)$ of resonator Nº 5: comparing measured and simulated data
Besides being shifted along the frequency axis, the peaks of the 1st resonance are less sharp in the FEM data and the phase does not breach far into the negative range as seen in the measured data. This mirrors the fact that the Y- and Z-circles (fig. Q.11) are smaller and due to similar or even larger vertical offsets much less centred on the real axis. The data of the lab recordings shows that the first resonance can also be interpreted as an immutable wide structure which is in some cases overlaid by a narrow additional structure. In the context of fitting equivalent circuit models, this is the argument for comparing not only the two first resonances of the cyan and the green curves and adding an additional comparison, where in the case of the FEM data, the Y-circle of the left is combined with the Z-circle of the right peak of the split first resonance.

although switched. A fact that the lab and the FEM data have in common is that the mechanical Q-factors determined from the electrical properties are 20-30 % lower than the values deduced by the sharpness of the sound pressure peaks. The coupling factors k with ≈ 0.2 and 0.05 for the first two resonances are similar in the three tables (O.4 O.5 & Q.5). Another feature not in agreement is the Z-circle size, where a huge difference can be seen for the first resonance in figure Q.11. Remembering the aspects described in appendix J.5, this hints towards issues of dielectric damping and parallel capacitances. Finally, it is interesting that the measured Y-circles within fig. O.22 and within fig. Q.11 seem to have similar vertical offsets, whereas the group of FEM data-based Y-circles in figure Q.11 shows a substantial variation in that respect. The question about the particular FE model features which influence this was not further studied.

The picture one can see when looking at the SF resonator and its FE model from the electrical side could be summed up as follows: while the examined resonator is a complex system showing complex pressure map and wall displacement patterns rich in structure, and while the re-appearance of key elements of these structures in the FEM simulation allows the conclusion that essential features are captured (although shifted), the electrical data is much more simple in comparison. It is too poor in structure to allow the easy identification of corresponding detail

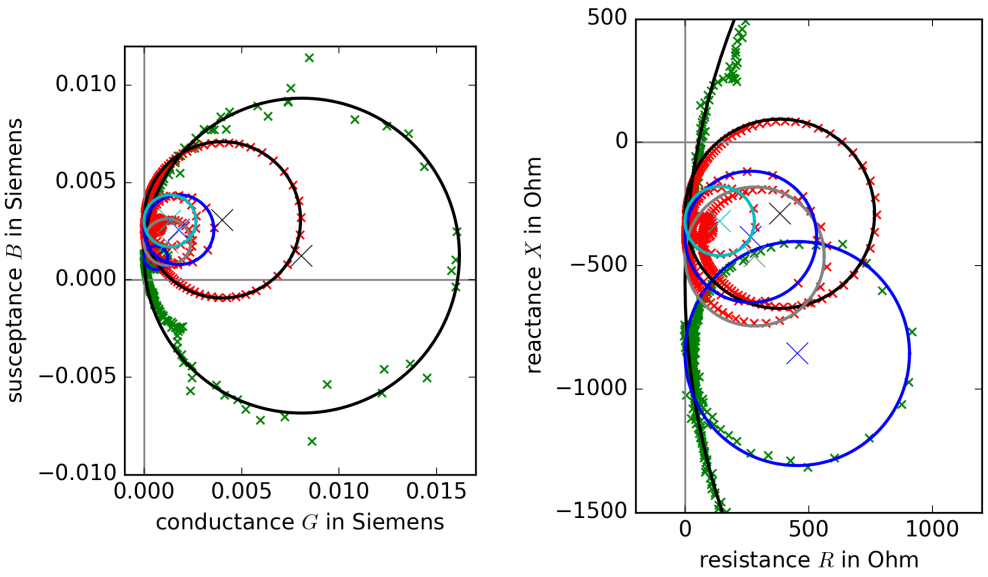


Figure Q.11 Y- and Z-circles of resonator № 5: comparing measured and simulated data

In these two plots of the complex admittance (left) and impedance (right) planes the measured data (dataset 129, shown in green) is overlaid by the data read from the FEM simulation results (red). In order to distinguish the different resonances, the fit circles which were created during the Y- and Z-circle analyses are added to the plots in different colours: the 1st, 2nd, and 3rd resonance are plotted in black, blue, and cyan, respectively. In the lab measurement, only the first two are covered. In the FEM data, as the 1st resonance is split, there is an additional circle plotted in grey, which corresponds to the second peak of the 1st resonance.

features at this point. (Displacement and pressure modes can be distinguished and paired by their profiles, but Y- and Z-circles look all the same except for the size.) However, the electrical data might eventually become more useful as a very clearly interpretable and quite sensitive model error measure in the next future iteration of resonator characterisation and FE modelling efforts when the gap between reality and simulation should anyway be a lot closer.

Q.2.5 Mesh dependencies

Influence of mesh shape

Although in FEM simulations of structural mechanics it is much less of an issue than in e. g. finite volume (FV) methods applied to CFD problems, the difference between irregular and regular mesh versions was examined. The APDL scripting language offers various tools for covering a 2D geometry piecewise with regular (mapped) meshes of triangles or quadrilaterals. The basic idea is that as rectangles can be map-meshed, so can be parallelograms; and as triangles can be map-meshed, so can be quarter circles or other circle segments. Similarly, many areas can be covered by a mapped mesh, even if they possess more than three or four corner points, if the bounding line can be divided into three or four suitable subdivisions, so the mesh of a triangle or rectangle can be stretched into the same shape. Figure Q.12 shows the two versions of mapped meshes created for the FE model of resonator 5

by subdividing the geometry into suitable three- and four-sided area patches. The exact sequence of line divisions becomes important in creating error-free meshing routines. As for opposing faces the element count must match, always the wider face needs to be discretised first in order to not breach the element size limit on the other side. This leads to fan-like structures of narrowing lines, to element aspect ratios substantially deviating from unity, and generally to more elements needed in the mapped versions as compared to free meshes. In the patterns shown in figure Q.12, the amounts of elements (and nodes) needed are 6300, 8000, and 8500 (10 000, 12 000, and 12 500).

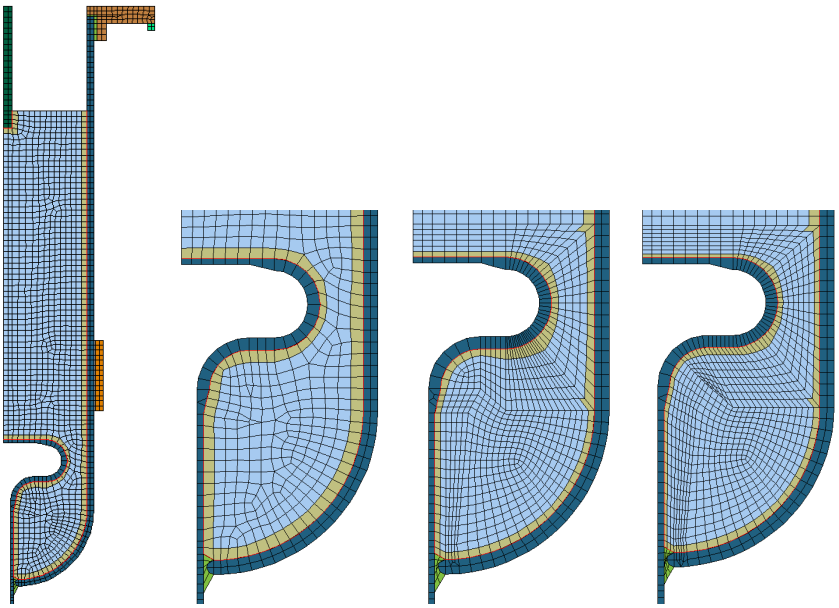


Figure Q.12 Mapped and free FE mesh alternatives
Three meshing routines were written for an early version of the resonator 5 geometry: one for a free mesh and two slightly differing versions generating mapped meshes. The second mapped mesh version, shown all the way to the right, aims for a less inhomogeneous element size distribution. For the above plots the element size parameter was raised to 1.5 mm to allow better visibility. The difference between this geometry and the later one shown in figure Q.5 is the shape of the round bottom of the main glass part. Here this is a simple circle segment, whereas in fig. Q.5 this part is vertically stretched into an elliptical shape, which lets the rounding begin further up and nearer to the transducer and makes it more similar to the real-world counterpart. No mapped mesh generation routine was written for the later version because for the purpose of the mesh size study both geometries are considered to be of equal value, implying no need for a repeated mesh size study on the updated geometry.

An interesting comparison of FEM simulation results gained with the three different meshes can be found in figure Q.13. The plot shows the sound pressure response picked up at two points, the hydrophone tip at 3.5 cm (upper plot) and at a node on the central axis 2 cm below (lower plot). The colours in each triplet of curves, ranging from darker to brighter, correspond to the mesh versions, from left to right, in figure Q.12. The mesh resolution levels are determined by the APDL command `aesize`, which sets the limit for the longest allowable element edge and serves as the criterion for the number of line divisions made on the area boundaries. The lowest groups of curves correspond to the smallest `aesize` setting of 1 mm. The upwards shifted triplets represent settings of 2 and 3 mm. To read the data as a

general indication of a worse performance by the free mesh would be an oversimplification. What the odd curves in the upper two groups show is that the quality of the results obtained with the free mesh degrade much more quickly, and secondly, that there is a transition from small deformations of the pressure frequency response to drastic changes of the whole landscape. As is visible in figure Q.12, the element size distribution is less homogeneous in the case of the mapped meshes where narrowing lines exist in fan-shaped structures. The resulting areas containing element sizes far below the given limit may be contributing to this behaviour because it is possible that the large errors arise locally. For one thing, the low impact of the mesh variations on the second resonance, the one mainly dominated by the fluid volume, proves that this region does not become underresolved in any of the settings, which is also expected because the largest element size of 3 mm still lets 20 elements fit into the volume diameter of 6 cm representing half a wavelength of the sound field. But the wavelengths of the displacement patterns in the structural parts are on a much smaller scale. Thus one can assume that structural parts become underresolved first. Furthermore, one can imagine that not all structure parts are of the same relevance to the vibration mode shapes and eigenfrequencies, but that some key areas of largest degrees of bending[10] have an elevated impact. The conclusion is that one particular geometry detail happens to be the first one to become severely underresolved before other regions do. It is assumed that this threshold is only reached in the freely meshed FE model, but not in the mapped versions. Therefore, it is debatable whether figure Q.13 really shows the influence of the mesh type, or not rather the effects of local mesh resolution variations. In any case, it becomes clear that a mesh resolution study should not only be based on examining the second resonance, although it is the working point of interest of the resonator, but that one of the other modes should be included as well.

Influence of mesh resolution

Results of FEM studies are only meaningful, if it can be proven that the errors introduced by the approximations of the numerical concept are controllable and sufficiently small. Mesh size studies are a common practice to judge the influence of the fundamental approximation underlying the FE method, the spatial discretisation, because in principle the error should disappear in the limit of an infinitesimally small grid spacing.

For tracking the solution quality over the element size, physically meaningful integral solution quantities should be chosen as the observables. Plotting the observables over the element size should then reveal either asymptotical or a power law approach towards an aiming point, ideally the true limit value. The ensueing decision on the mesh resolution is always a trade-off between cost and accuracy.

[10] In the case of a steel ball on a steel spring, the Young's modulus of the ball does not influence the eigenfrequency, only its mass does. Similarly, in an oscillating beam with one clamped and one free end (ruler on table rim), the free end could be made of a softer or harder material of similar density, and the influence on the vibration modes would be little. Because of the low stresses and deformation levels near the free end, it may also be underresolved in an FEM simulation without posing a big problem.

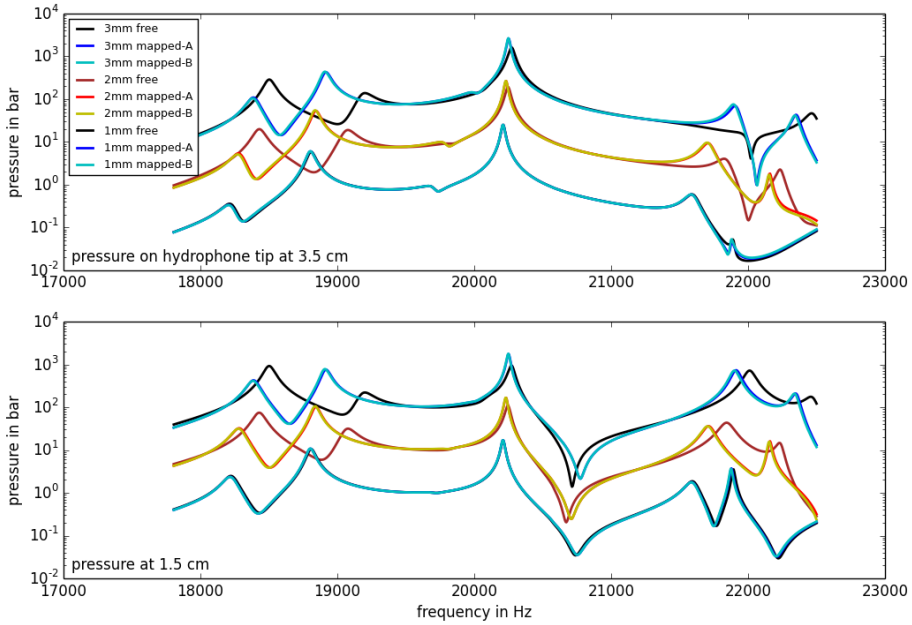


Figure Q.13 Influence of the mesh type for three mesh resolution levels
Nine simulations are compared based on the combination of the three mesh types with three element size settings. The upper plot shows the sound pressure on the hydrophone tip, located at $y = 3.5\,\mathrm{cm}$ in this case. Since for the 3$^{\mathrm{rd}}$ resonance this position corresponds to a sound pressure node, the lower plot is added, showing the sound pressure measured at the FE node on the central axis nearest to $y = 1.5\,\mathrm{cm}$. The lowest group of curves in each plot shows the results with the finest mesh resolution where the element size limit is 1 mm. The datasets generated with coarser meshes (limits of 2 and 3 mm) are shifted upwards in the logarithmic plot by multiplication with 10 and 100.

Observables

The hydrophone signal and probed glass wall displacements were up to now described as the principal quantities for judging the FE model quality in a comparison with available experimental benchmark data. But are they also suitable for judging the quality decay of the simulation due to a lowering of the mesh resolution? Not really. The reason is that these quantities are read out locally, they stem from individual nodes of the FE mesh which have variable neighbourhoods or can change their own position. In the case of a free mesh even the local topology can change. The hydrophone signal is computed as the mean pressure loaded upon the nodes forming the hydrophone tip surface. But still, the number of these nodes is very small and each additional node entering the set during refinement can have a discontinuous impact on the signal. Therefore, it is preferable to base the element size study on integral quantities representing observables of the entire FE model. The two integral and physically meaningful observables chosen are the average pressure amplitude seen by the entire set of nodes of the acoustic fluid and the average of the local stress intensity seen by the entire set of nodes of the structural materials. The local stress intensity σ_I is defined as the largest difference between principle stresses σ_i,

i. e.

$$\sigma_I = \max\left(|\sigma_1 - \sigma_2|, |\sigma_2 - \sigma_3|, |\sigma_3 - \sigma_1|, \right). \tag{Q.4}$$

As the shape of a given oscillation mode is a characteristic property of the whole system, it is possible to add two more observable quantities to the set, the exact resonance frequency and the amplitude of a mode. This can be done if it is ensured that these numbers are not generated from reading out single nodes. A very elegant method would be to decompose the sound field (excluding the peripheric areas near the structure) into spherical wave functions and use the first coefficient of the series [190]. A much simpler approach is to fit a cosine function to the central part of an antinode of a sound pressure mode shape. The latter approach was taken here.

Problem of shifting resonance peaks

There is one problematic aspect: changing the element size does not only change the amplitudes of resonance peaks, it also affects their positions on the frequency axis. (This is because a different mesh means a different model.) Consequently, simply choosing the frequency of a resonance peak f_{res} and examining what happens to the observables taken again at that frequency f_{res} after a change of the mesh resolution, may lead to mistakes in interpretation. A 10 % change in the observable after a mesh coarsening does not necessarily mean a substantial worsening of the model quality. In the case of a sharp peak it may merely be the consequence of a tiny frequency shift of the resonance. To avoid the trap, one option is to examine a frequency far away from any resonance. A second option is to follow the resonance and track the change of both, the observales' magnitudes and the resonance frequency. The drawback of the first is that avoiding peak oscillation amplitudes means at the same time avoiding the regimes where linearisations in the model equations have their greatest impact. The disadvantage of the second option is the increased computational cost of a sufficiently fine frequency resolution. The second option was pursued here while mitigating the cost increase through scripted iterative steps of zooming in on the frequency axis.

Results

The results of the mesh refinement study targeting on the second peak of the first resonance are presented in figure Q.14 and an analogue examination of the second resonance can be found in fig. Q.15. All three mesh types, the free and the two mapped versions were examined, and always on a resonance peak. The exact resonance frequencies were sought by going through four different zoom levels on the frequency axis. In the first zoom step the interval from 18.7 to 19.5 kHz was screened in 40 steps, and in the next three steps the intervals were of widths of 40, 4, and 0.4 Hz, centred around the resonance frequencies detected on the previous zoom level, and probed with the same amount of steps.

Three main features can be discovered from the plots: large shifts of the values in the cases with the one or two coarsest meshes, asymptotical behaviour for the finest meshes, and some degree of random scattering of the points. The interpretation for the large deviations, accompanied also by substantial shifts of the peaks on the

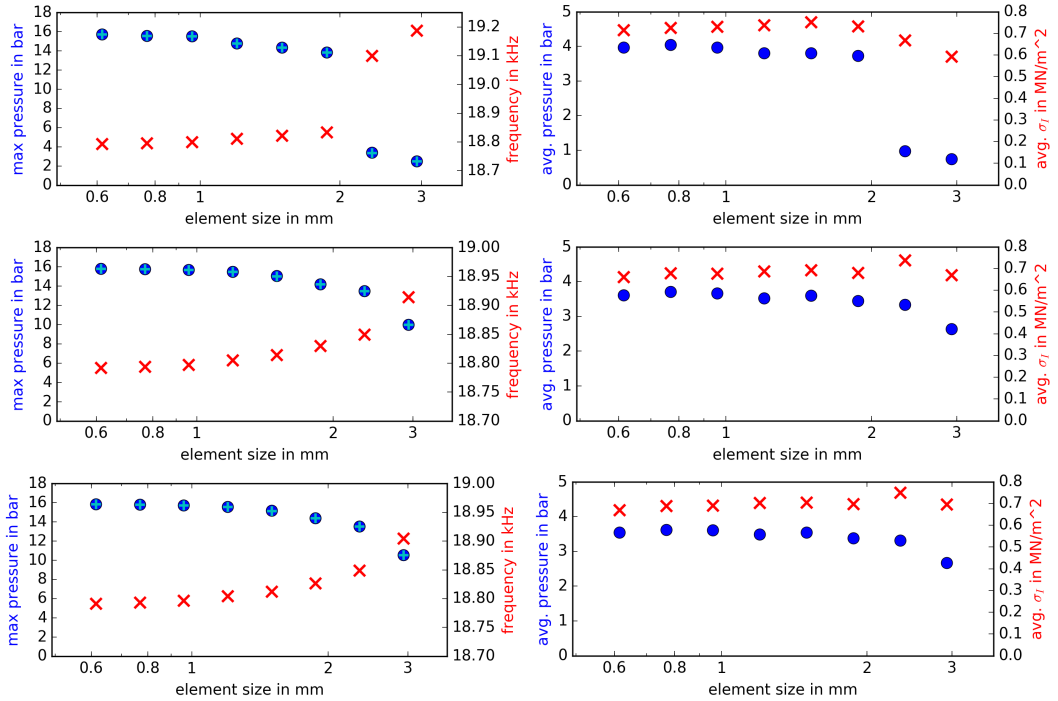


Figure Q.14 Varying the element size: impact on 1st resonance
These plots show how integral characteristics of the simulation results change under variation of the mesh resolution. The element size limit was set to $1.5\,\text{mm} \times 0.8^i$ with $i = -3, -2, ..., 5$ yielding values from 2.9 to 0.6 mm. With the free mesh version this created model sizes from 860 to 15 700 elements, with the mapped meshes the count ranged from 1100 to 21 100. The plots on the left side show the result of tracking the main pressure antinode in terms of amplitude and frequency, the plots on the right give the values of the two integral quantities, the average pressure amplitude in the fluid and the average stress intensity in the structural materials. The rows from top to bottom correspond to the mesh versions from left to right in figure Q.12. The left plots contain two layers of symbols for the pressure amplitude: the blue ones stem from reading out the one node where the highest pressure is measured, the cyan ones correspond to the peaks of cosine functions fitted to the nodal pressure values. For the fit only that segment along the central axis was considered where the amplitude was above 90 % of the peak value.

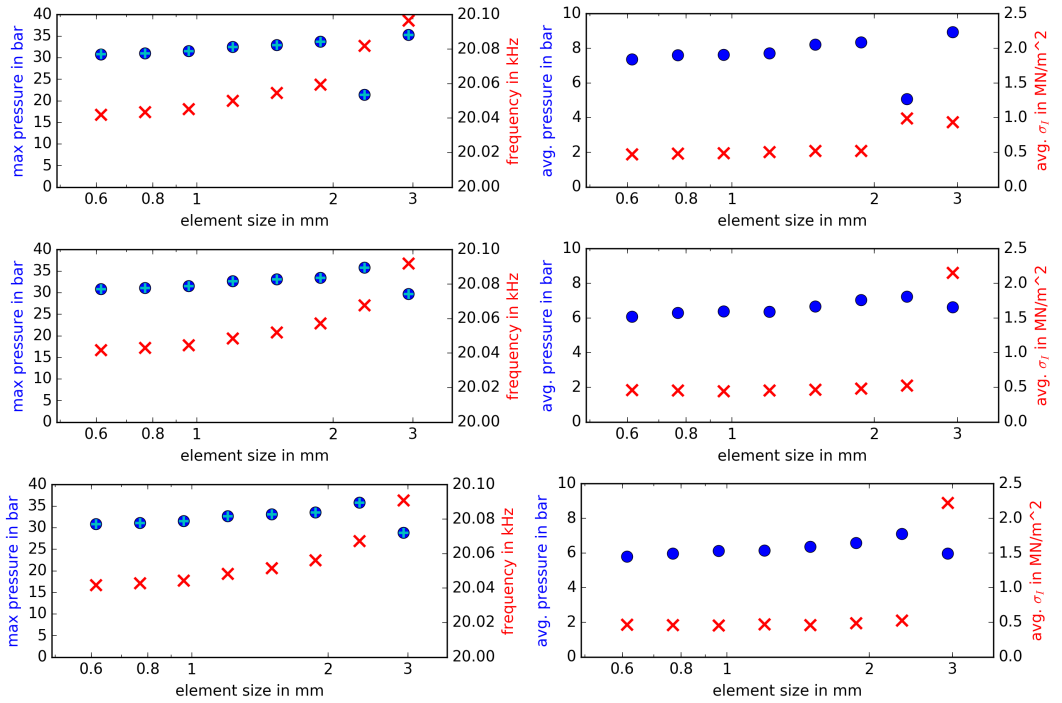


Figure Q.15 Varying the element size: impact on 2nd resonance
These plots show similar data as described in figure Q.14, but the target here is the second resonance, which is influenced less strongly by the applied changes in the mesh resolution.

frequency axis, which corresponds to the changes of the whole frequency response landscapes seen in figure Q.13, is that decisive local geometry details have become sufficiently underresolved to not be able to capture the physics correctly any more. The second feature, the asymptotic tendency for ever finer meshes, is the expected behaviour of the error introduced by the spatial discretisation. For the last feature, the slight scattering seeming like a small addition of randomness, a first explanation attempt may be to attribute it to a sort of sampling error, i. e. how the distributions of nodes sample the stress and pressure fields. But this can be excluded because it does not explain why the upwards and downwards deviations are in sync in all three data types. Therefore it is speculated that it may be the consequence of how different node distributions allow the model more or less easily to fall into its vibration eigenmode.

As concerns the central question any mesh resolution study should answer, the question about the element size limit beyond which the simulation becomes unreliable, the conclusions can be drawn that below 2 mm the problem of wrongly simulated physics due to local severe underresolution disappears, and that after going down to 1 mm also the discretisation errors have been minimised by a large degree. It is to be seen in that context that the simulations of resonator Nº 5 discussed above were conducted with an element size limit of 1 mm, while the evolutionary optimisations of new designs presented in chapter 5 were done with a setting of 1.5 mm.

Using experiments with local mesh refinements in order to learn more about the

types of geometry details prone to becoming severely underresolved first, could make sense for speeding up future optimisation runs as it might allow making large parts of the meshes coarser.

Q.3 SF resonator sensitivity study

It is a central point of interest here to ask the question of how difficult it is to reproduce a sonofusion resonator of the West-Howlett design as used for SF experiments by Taleyarkhan et al., i. e. how reliably the same sound pressure performance of one resonator can be achieved when trying to replicate it. During the experimental studies at RPI, it was noticed that different exemplars of the same resonator design yielded different sound pressures, and single exemplars showed shifted resonance frequencies with slightly varying sound pressure from day to day (drifting properties). It was assumed, as pointed out in appendix chapter I.4, that these variations originate i. a. from changed liquid filling levels, piston positions, or slightly differing glass part shapes. The present sensitivity study[11] was conducted in order to check whether these assumed explanations for the observed resonator performance variations are valid.

Q.3.1 The model geometry and parametrisation

The parameter study aims at exploring the sensitivity of the SF experiment setup, this means a closed resonator geometry needs to be simulated. As the characterised RPI resonator is equipped with the hard to simulate aluminium flanges, and as only the ORNL and Purdue setups with the silicone bead connections instead of the flanges are said to have yielded neutron counts, the geometry choice is straightforward and falls on the one depicted in Taleyarkhan's sketch (app. D), i. e. the West-Howlett resonator with a top head sealed by silicone bead. The meshed geometry is depicted in figure Q.16. The material properties used in this model involve the preliminary set of piezoelectric constants (see table Q.1) and also a preliminary set of loss tangents: 0.003 for glass, 0.1 for silicone, 0.001 for the fixation material, 0.001 for the piezo ceramic, and 0.01 for its dielectric loss tangent. A conference paper [435] documents the deduction of these loss tangent values through calibrating the model against measurements. The next question is about the geometry details, i. e. the exact combination of design parameters furnishing the reference dataset for the parameter study. Which should be the central anchor point of the star-like variations in the parameter space along each axis? Simply abiding as strictly as possible to Taleyarkhan's sketch and descriptions leads to an FE model that does not reveal the strongest sound pressure performance. This means picking any design parameter and incrementally modifying it does not only lead to deteriorations of the sound pressure, some changes improve the pressure field. The question should be asked whether the parameter study would be equally telling with either a mediocre

[11] The sensitivity study was published as a conference paper [437]. In an older paper on the first insights gained with the newly established FE model [435] the research status was described which motivated the parameter study.

or a rather good reference set serving as the anchor point. It is assumed that exploring the vicinity of a good setup brings more insights. Why that? Assuming that a lot of stiffnesses, distances, and masses need to fit together for enabling a good working mode of the resonator, it seems more probable that regions of high performance are slim insular peaks or narrow ridgelines in parameter space rather than plateaus. Then, it is clear that a sensitivity study somewhere far away from any performance peak or ridgeline gives no insight on the difficulty level to reproduce well-performing resonators. Consequently, a pre-optimised geometry was sought by tuning several strongly influential design parameters. The design variations of the RPI resonators show that parameters like the inner chamber height, the vertical positioning of the transducer, and the piston diameter were the target of experimentation already before. After a procedure of tuning these parameters manually aiming at maximising the amplitude of the second resonance the geometry depicted in figure Q.16 was reached. The parametrisation of the model used for the tuning as well as the sensitivity study is shown in figure Q.17 and further described in table Q.6.

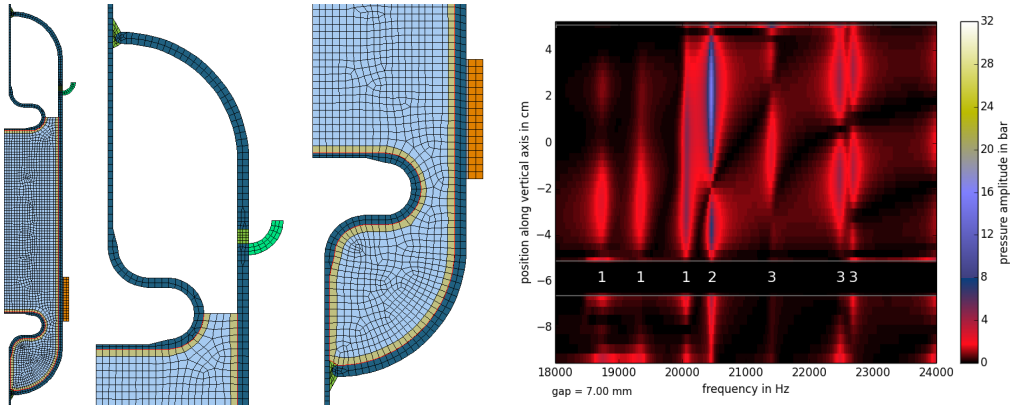


Figure Q.16 Base geometry of the SF resonator sensitivity study
This FE mesh was used to simulate the acoustic properties of the type of resonator used for the controversial SF trials of Taleyarkhan et al. [458]. This geometry was modelled after Taleyarkhan's sketch (see app. D) and the RPI replicas. Then, several key parameters (i. a. inner height, transducer position, piston diameter) were tuned manually aiming at the strongest possible amplitude of the second resonance. Only a pre-optimised resonator is deemed to be a good start for a meaningful sensitivity study. The resulting setup is depicted above. The sound pressure response featuring the dominant 2^{nd} resonance is added on the right.

This pre-optimised geometry was taken as the reference point for exploring the design parameter sensitivity of the resonator's sound pressure performance. The pre-optimisation procedure could build on the available collection of pre-existing design variations of this FEM geometry. Its final part consisted in tuning the three highly influential design parameters (inner height: *ih*, transducer position: *pipos*, piston diameter: *gap*) in two steps by first scanning a 2D space $ih \times gap$ and afterwards $ih \times pipos$ in 7×7 steps and each time picking the best combination to continue.

Looking at the displacement map and the deformed shape corresponding to the strong 2^{nd} resonance, both presented in figure Q.18, we can discover again what has already been revealed by the simulation of resonator 5 (and which is visible in fig. Q.9), that the radial displacement of the glass wall a few centimetres above the transducer is more than twice as large as the displacement directly at the trans-

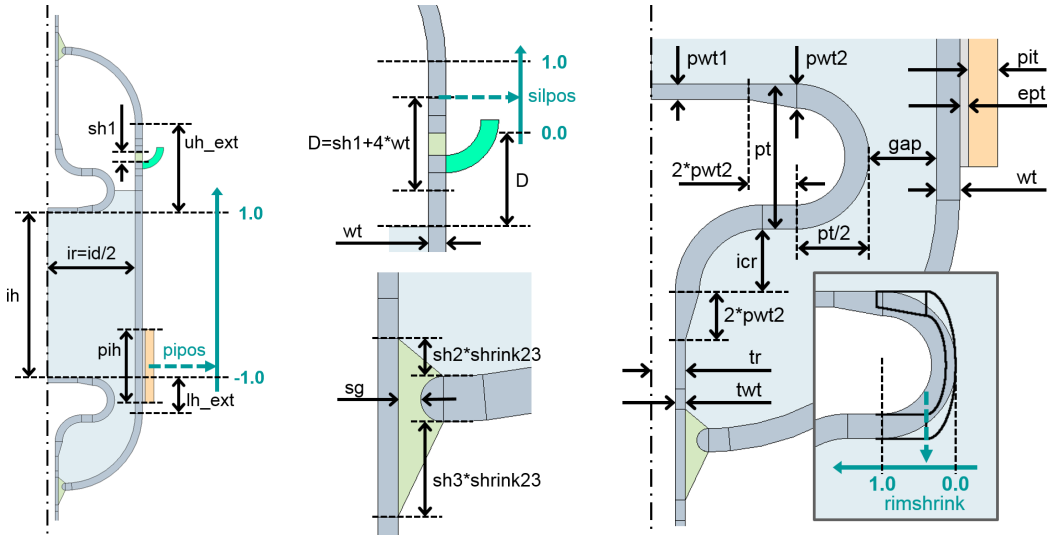


Figure Q.17 Geometry parametrisation
The geometry parametrisation utilised for the sensitivity study is depicted. Parameters given in absolute numbers are labelled with black signs, for scaling or relative translation parameters there are extra coordinate systems drawn in green. Concerning the two pistons, the parameters apply to the upper and lower piston equally, the case of two differently shaped pistons was not considered.

Table Q.6 Parameters of the sensitivity study
This table lists and explains the parameters that were varied during the sensitivity study. The last two columns indicate the nominal value of each parameter and the variation step size δ used for composing the bar chart in figure Q.21. The unit indicated for the last column does not apply to the dimensionless scaling parameters *pipos, rimshrink, shrink23,* and *silpos*.

label	description	nominal [mm]	δ [mm]
ih	inner height (piston-to-piston distance)	102	2
pipos	transducer position, −1 is lowest, +1 is highest	−0.86	0.04
gap	horizontal gap between piston rim and inside of main wall	7	0.5
pwt1	piston wall thickness 1 (front plate)	1	0.2
pwt2	piston wall thickness 2	2	0.2
pt	piston thickness	15	1
rimshrink	horizontally stretch piston rim while gap=const.	1	0.1
icr	inner curvature radius of piston-tube connection	6.5	1
ext_lh	lower half extension; distance from reflector (piston front plate) to base of main wall rounding	20	2.5
ext_uh	upper half extension; distance from reflector (piston front plate) to base of main wall rounding	40	2.5
sg	width of silicone gap for piston tube fixation	0.5	0.005-0.1
shrink23	vertically shrinking or expanding silicone bead around piston tube	1	0.25
sh1	height of silicone bead connecting main cylinder and top head	3	0.2-1
silpos	position of silicone bead connecting to top head; [0,1]=[low,high]	0.5	0.1
tr	piston tubes radius	1	0.2

ducer. This means the glass wall moves in a manner which amplifies the transducer displacement, and the large amplitude of the upper antinode of the sound pressure profile of the 2nd resonance is related to this feature.

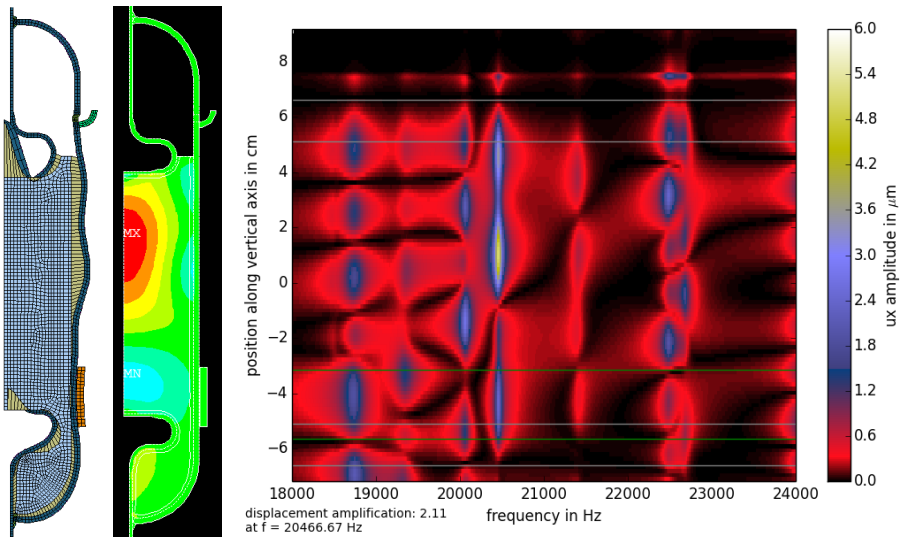


Figure Q.18 Displacement field and pressure snapshot of the pre-optimised resonator

The two cross section plots on the left show a deformation snapshot and a pressure field snapshot for the second resonance near 20.5 kHz. The diagram on the right is a displacement amplitude colour map indicating the normal displacement along the inside surface of the outer glass wall. A close look at the two displacement plots reveals that for the second resonance the strong sound pressure amplitude of the pre-optimised setup is enabled by a displacement amplification mechanism along the glass wall: the upper part of the glass wall neighbouring the strong upper central sound pressure antinode moves more than twice as much as the glass wall a few centimetres below where the transducer sits; the factor is 2.1. The 20.5 kHz resonance is special when comparing it with the wall displacement patterns at lower or higher frequencies due to the much lower spatial wavelength. It may be interpreted as the fluid forcing its motion pattern onto the glass wall.

The sensitivity study was conducted in two steps: on the one hand the three geometry parameters *ih*, *pipos*, and *gap* were varied over larger value ranges and on the other hand a much larger set of design parameters was varied in a narrow range, three small steps upwards and downwards. The impact on the sound pressure amplitude are visualised in figures Q.19, Q.20, and Q.21. These figures transport an important message because they are the clearest illustration of how sensitive this resonator design is, of how precisely one has to hit a design point in order not to miss an envisaged peak of sound pressure performance. Thus, they illuminate the conflict of the low tolerances required by the detected level of sensitivity and the larger existing tolerances due to the manufacturing process involving e. g. manual glassblowing work. At the same time these three figures furnish a telling characterisation of the type of search landscape one has to deal with when intending to apply an algorithmic optimiser to the design problem.

The question of how to investigate and visualise the sensitivity of the sound pressure performance is not completely trivial. Only keeping an eye on one single initially chosen resonance will give an overly pessimistic result because if the tracked resonance decreases drastically in amplitude there might be other resonances developing nearby offering a better working point. Not tracking any resonance and simply

measuring the strongest sound pressure amplitude occurring in the chosen frequency band will yield an overly optimistic result; one might easily derive an assuasive statement that the probability is low that all resonances in the scanned range are weak, but it would be a useless or even misleading statement as it neglects that not all resonances offer equally suitable working points for generating violently bursting and collapsing cavitation bubble clusters.

Thus, when comparing different design variations, it is necessary to make judgements about the different emerging resonances. As judgement in the form of direct deliberative picking among several available resonances would open the door for subjectivity, a much better alternative is to program a postprocessing routine for resonance characterisation which treats all produced design variation outputs equally. This is particularly necessary as the closed resonator FE model generally shows a richer resonance structure as compared to the open chamber model which was used for benchmarking, as can be seen from comparing figures Q.7, Q.8, and Q.19.

Therefore, a routine was programmed with the initial aim of classifying the many occurring resonances into types “1”, “2”, and “3” according to the profiles in between the pistons having one, two, or three antinodes. Due to constraints of practicality and in order not to neglect profiles outside the clear pattern the routine was made to detect type-2 resonances by identifying a two-belly profile with the higher amplitude at the upper antinode and to simply sort all other resonances into one group at lower and another one at higher frequencies independently of the profiles. Examining the maximal pressure peaks found within these three groups offers more information than only counting the overall pressure maxima within each map while not neglecting any resonance due to unsystematic picking. Figure Q.19 offers examples of how the resonances were ordered into the three groups. For determining the pressure peak of a resonance, only the segment of the central axis profile in between the pistons was considered. This segment was further truncated by 3 mm at each end in order to judge a resonance only by the pressure amplitude it produces in the liquid bulk and ignore antinodes attached to the piston front plates.

A quite important detail which can be observed in the pressure map variations compiled in figure Q.19 is the *avoided crossing* (or *anticrossing*) of resonances. Anticrossing is the direct consequence of a coupling force between two oscillators [329]. In this context the acoustic resonator can be imagined as a set of coupled oscillators because the piston, the outer hull, the liquid, these sub-parts have internal vibration modes and the assembled resonator has emergent complex vibration patterns. In the depicted series of pressure maps the variation of the piston diameter induces the type-1 resonance located in the first plot at 20 kHz to move to higher frequencies. In the fifth plot it approaches its neighbouring type-2 resonance very closely, it turns itself into a type-2 resonance, and finally the old type-2 resonance, as if being pushed away, starts moving to the right. The two resonances never merge into one. Most significantly, the amplitude of both resonances is low at the moment when their distance is closest. In order to understand the task of optimal resonator design, instead of a superposition of resonances it is therefore better to describe the situation as a competition of interacting resonances.

The pressure maps of figure Q.19 show the classification of resonances, their changing amplitudes and positions on the frequency axis, and they indicate whether

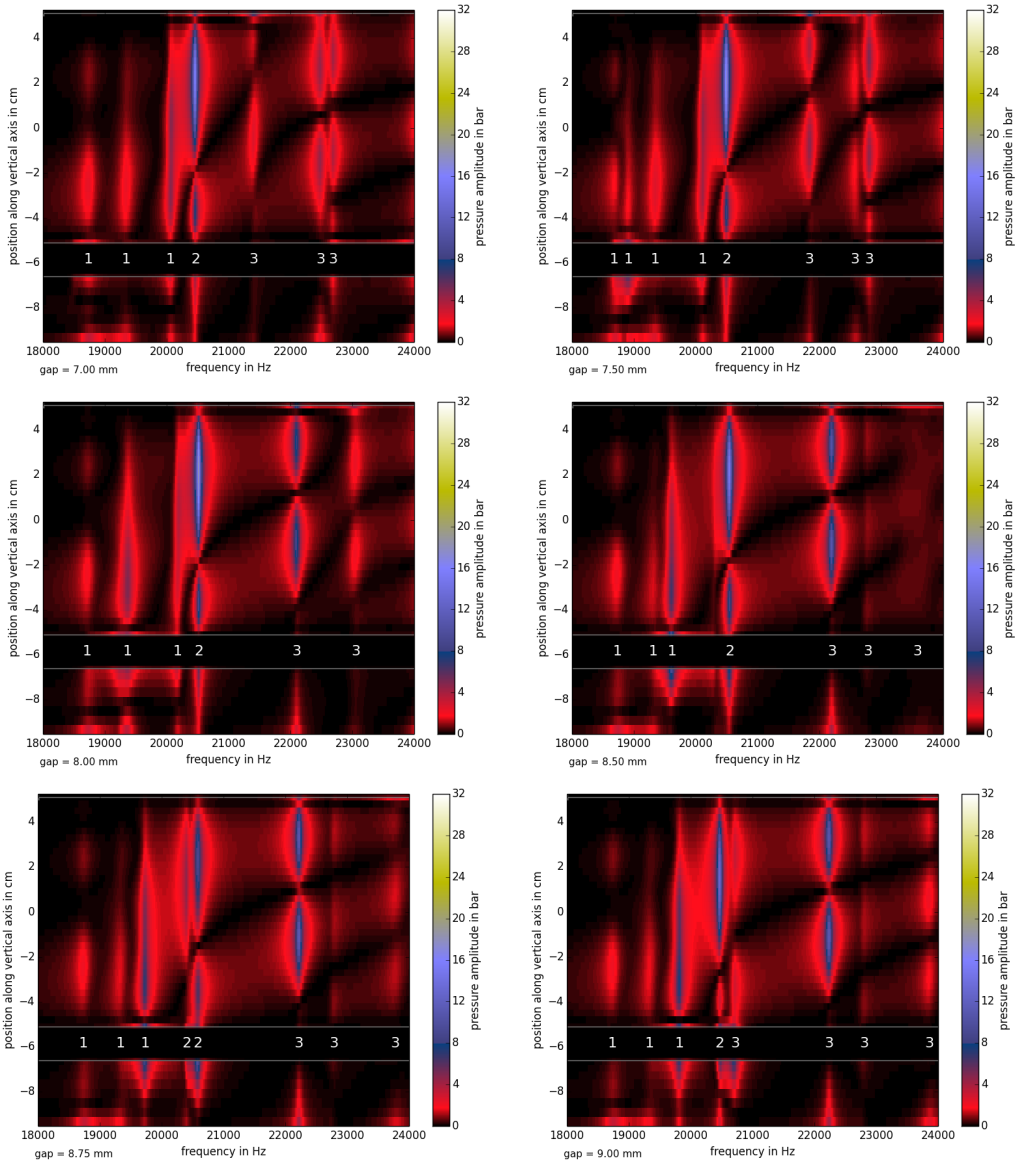


Figure Q.19 Resonator sensitivity due to competition of oscillation modes
This collection of pressure amplitude maps illustrates the effect of varying the parameter *gap* in five steps of half a millimetre. The series shows that some structures move along the frequency axis while others stay in place. The interaction of moving resonances has a major impact on peak amplitudes, as is exemplified by the substantially lowered amplitudes of the type-2 resonances in the fifth plot ($gap = 8.75\,\mathrm{mm}$). A relevant detail visible from the six different expressions of the 2^{nd} resonance is how different boundary conditions emerge on the front surfaces of the top and bottom pistons: sometimes there is an almost perfect pressure node while in other situations a substantial pressure amplitude builds up on a piston surface. In order to mask such pressure peaks on the piston surfaces the profile segments in between the pistons were truncated by 3 mm at each end when determining the peak amplitudes.

at the centres of the piston front plates there is a pressure node or not. The results of postprocessing many such pressure maps can be compiled to give a better view of the sensitivity of the resonator. Besides the pressure maps from inside the liquid along the central axis, it is pressure maps from data gathered along the entire fluid-structure interface which have to be taken into account to compute a useful value of the wall pressure ratio, as the peak interface pressure amplitude can occur at many different locations.[12] The result of analysing 79 design variations is depicted in figure Q.20. It shows how the pressure peaks rise and fall, how the frequencies shift, and how the wall pressure ratio changes as a consequence of tuning the inner height, the transducer position, and the piston diameter, all based on the automatic classification into three groups of resonances. It illustrates that regions of good sound pressure performance are broken up by regions of deep pressure drops. Some of these sharp gaps can be attributed to anticrossing events quenching the amplitudes. It is of central importance to note that many of the design variations with strong pressure performance are disadvantaged by elevated wall pressure ratios. Only few working points have a wall pressure ratio below 40 %. This makes it very clear that a well-proportioned resonator of the West-Howlett design will make SF experiments much more promising than blind trials based on approximative knowledge of suitable design parameters.

Does the situation indicated by figure Q.20 also mean that finding good proportions is a really challenging optimisation problem? Not necessarily. If besides the three investigated parameters all others were to have a weak effect on the resonator performance, then the search for the right setting in a 3D space would be safely solvable based on systematic parameter variations of the FEM simulation.

A second series of design variations was conducted to answer whether the resonator performance is very sensitive only with respect to a few parameters or to a larger number. A larger selection of design parameters of the FE geometry were varied by three incremental steps upwards and downwards. This selection includes parameters determining the shape of glass parts and silicone connections, properties which in the real world are the result of manual glassblowing and assembly work. A few parameters like the radius of the main glass cylinder which is an outcome of a much better controllable industrial production process were not included (available glass tube sizes are the given starting point for the resonator design). The variation increments (listed in table Q.6) were deliberately chosen for each parameter and based on the knowledge of the manufacturing process and the shapes being present in the RPI collection of resonator exemplars. The peak pressure data of that series are compiled in the form of a bar chart in figure Q.21. That chart shows that only five of fifteen parameters have a weak effect of less than 5 % pressure de- or increase, all other parameters have effects larger than 20 %. Four parameters breach the 20 % threshold already with a single increment step.

One parameter is included even though it is the result of industrial manufac-

[12] As mentioned in previous chapters, a high wall pressure ratio, i. e. the ratio between the peak amplitude in the liquid bulk and the peak amplitude found along the fluid-structure interface, will lead to cavitation on glass and silicone (and metal) surfaces and render a resonator unsuitable for SF experiments because this will deteriorate the pressure field. Generally, liquid-structure interfaces can bear less tension than the liquid bulk.

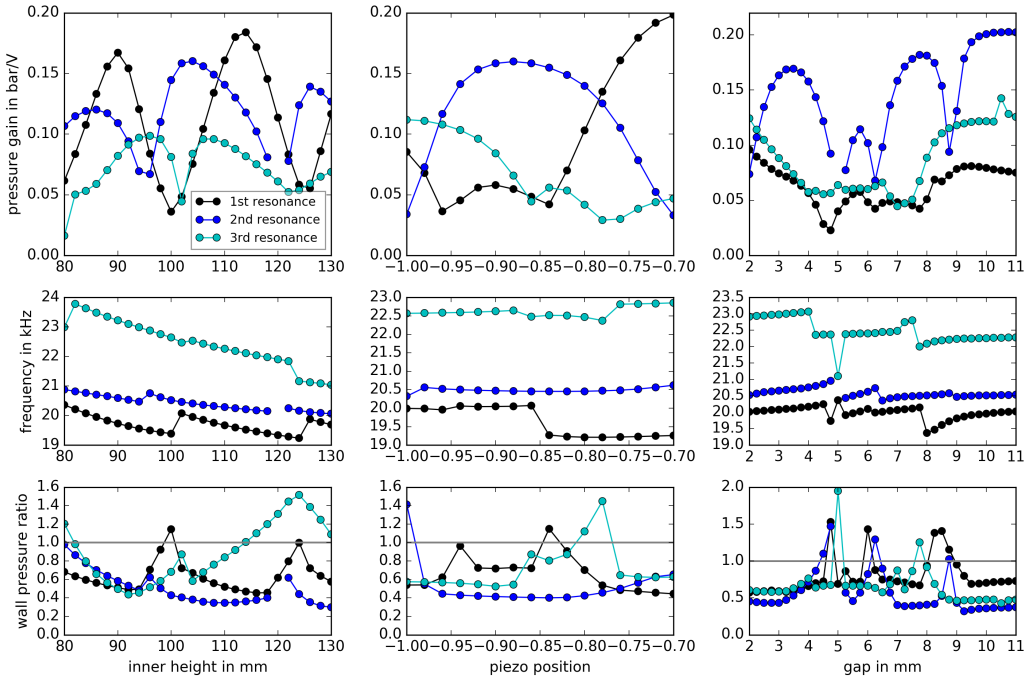


Figure Q.20 Sensitivity of the resonator model to three key design parameters
These plots illustrate the sensitivity of the resonator by tracking pressure peaks of several resonances and corresponding wall pressure ratios. The traces coloured in black, blue, and cyan represent the strongest 1st, 2nd, and 3rd resonance based on an automatic resonance group classification as described in the text. The frequency band from 18 to 24 kHz was investigated. The three plots of the top row show the variation of the amplitudes of the pressure peaks. The plots in the second and third row indicate the shifting resonance frequencies and the rising and falling wall pressure ratios. One of the important messages is that in order to design a well-performing SF resonator, the right parameter combination has to be chosen which maximises the amplitude peak of a resonance offering a suitable sound pressure field. The data shows that the requirement of a low wall pressure ratio can drastically reduce the number of suitable setups and working points. In the upper middle plot the wide arc representing the 2nd resonance indicates that for the transducer position the optimal region is wide enough so that there should be no problem with manufacturing precision. The plot to the left shows that for the inner height a 4 mm reduction from 100 to 96 mm can turn a good 2nd resonance profile bad. For the parameter *gap* the situation is most problematic, as two of the quarter millimetre steps can make the difference.

turing: the radius of piston tube. Just like the length of the tube, which is also controllable with high precision, it determines the elasticity and the mass of the tube. As the tube forms a counterpart mass for the piston front, its dimensioning is decisive for enabling a suitable internal vibration pattern, and therefore it is not surprising that the tube radius *tr* belongs into the group with large sensitivity effect. For parameters like this one the required precision should pose no challenge, but still each represents one more dimension enlarging the parameter space which has to be scanned when trying to systematically determine an optimal resonator setup.

To summarise the main message of the sensitivity bar chart: there is only a minority of parameters with weak effect, most of the parameters have a severe impact on the sound pressure performance. All those parameters have to match tightly, many with tolerances of fractions of a millimetre, if one wants to build a well-performing resonator of this setup. It is a reflection of the fact that many different masses, spring constants, and dimensions have to fit together.

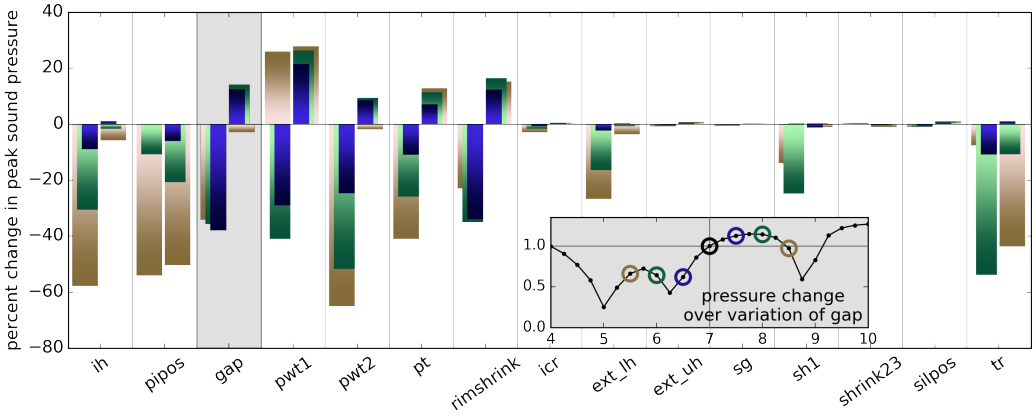


Figure Q.21 Sensitivity with respect to design parameters
Each parameter was varied three steps downwards and three upwards. The step sizes are given in table Q.6. The bars indicate the percentage of sound pressure amplitude change induced by the design variation. The three bars of the left group associated to each parameter represent the effect of downwards parameter variations and the bars of the right group the effect of upwards shifts. Blue, green, and brown bars represent the steps of one, two, and three increments, respectively. The inset illustrates the situation for the parameter *gap* by marking the values taken for the bar chart on the fine resolution variation sequence shown already in figure Q.20.

Another sensitivity test can be made by varying material constants instead of geometry dimensions. This was examined for the stiffnesses of glass, liquid, and piezoelectric ceramic. The FE model predicts changes in sound pressure up to 5 % when the constants are varied by 2 %. This is no strong impact but it isn't negligible either considering that the uncertainty on the stiffness matrix coefficients of piezoelectricity are 2-3 %[13].

Q.4 Comparison to FEM simulations of Taleyarkhan's Purdue group

Taleyarkhan's team also started efforts to simulate the resonator with finite elements using the Comsol® software package [12, 66, 495–498]. The simulations are also 2D-axis-symmetric and the piezoelectricity is taken into account. Several different resonator designs were simulated reflecting the different experimental setups used and also the different purposes. Not all setups are geared at sonofusion experiments. The other interest of the group is in the use of the resonators as particle detectors. The FE models of Wang et al. [496, 497] are practically the same geometries as the opened resonator 5 and the closed version described above. Indeed, the pressure profile of the 2nd resonance discussed above is in agreement with figure 4.1 in [496, 497]. The mechanical boundary conditions are probably the biggest difference between the two FE models. Wang et al. decided to translate the fact that the resonator is resting in its seat on the bottom surface of the transducer to

[13]Partially this is an unavoidable consequence of the polycrystal sintering process where the temperature time profile determines a solution within a goal conflict: a short cool-down can serve to keep the element distributions homogeneous, a slow cool-down will allow larger crystal lattice domains but entail the bad consequence that the doping elements have time to segregate [359]

a "roller" BC with no allowed vertical displacement. Furthermore the upper piston of the closed setup is held by its glass tube being rigidly clamped by a zero displacement BC a short distance behind the piston back plate. There is no connection between the upper piston and the rest of the chamber. The details of the connection of the lower piston to the rest seem to have undergone changes, e. g. from [498] to [497]. In summary, the Taleyarkhan group has explored more different geometries, many accompanied with point-wise benchmarking comparisons to selected lab data, whereas in the RPI-KIT project fewer resonators were examined experimentally, but the experimental data collection was done with finer resolution and more different signal types, benefiting the comparison between lab and FEM data.

Q.5 Discussion and insights

The described SF resonator FE model presents an important step forward. Its geometry was made richer in detail as earlier works and set up in a parametrised manner. Admittedly, the model is not yet able to produce sound pressure and displacement maps that match the measured laboratory data down to the detail level (and whether a detail-level match can be expected with the current system and its sensitivity is debatable), but the correct simulation of the vertical pressure profiles of the most important resonance modes became finally possible.[14] Examining different variants of the current resonator geometry with the FE model has led to a deeper understanding of what defines a performant resonator with respect to SF trials. The vibration mode of the 2nd resonance reveals that a mechanical displacement amplification mechanism by the vibrating glass wall is a key ingredient allowing high sound pressures in the liquid and low pressure amplitudes where liquid interfaces with structure. Another key characteristic is the ratio of pressure amplitudes within the central antinodes versus near the walls. From figures Q.19 and Q.20 one can learn that this ratio changes drastically from resonance to resonance and from setup to setup.

The parametric study of a pre-optimised setup reveals a high sensitivity of the sound pressure performance to small variations (down to the range of millimetre fractions) of geometry details. This offers an explanation for the difficulty of reproducibly attaining beneficial experimental conditions with stable cavitation rates as experienced by the RPI team during their efforts of replicating the SF experiment. It can explain Tessien's statement "It's a nightmare to run it, and it breaks" [485] if underperforming resonators have to be powered with much stronger transducers and very high voltages. On a more principal level it offers an explanation for the status of the sonofusion controversy. If the reproduction of a performant resonator is a game of luck for one research team (Taleyarkhan's team is said to have built dozens of resonators), what are the odds for a different team with only a drawing and a description to start with to hit the right spot in the design parameter space with much fewer trials and create a resonator where all important masses and stiffnesses match?

The question of sonofusion has not yet been examined by researchers in the

[14] At the same time and independent of Wang et al. [497].

right way because a suitable manner of doing it involves a reproducible experimental setup. This project has led to a large base collection of experimental characterisation data and FEM simulation data, which will be able to benefit any future work of resonator characterisation, benchmarking, and redesign. The data collection, by bringing together electrical, sound pressure, and radial displacement signals, and because of the fine resolution of the recordings, is very comprehensive and allows telling benchmarking comparisons.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
B	susceptance
C	capacity
$\mathbb{C}$	complex numbers
c	speed of sound
c_{ij}^{E}	elasticity tensor entries given at constant electric field
d	diameter
d_{ij}	piezoelectric strain tensor entries
E	Young's modulus
e_{ij}	piezoelectric stress tensor entries
f	frequency
G	conductance
g_{ij}	piezoelectric (g) constant (ratio of open-circuit field developed in response to stress) [353, 516]
h_{ij}	piezoelectric (h) constant (ratio of field developed under strain) [516]
I	current
i	unit value of imaginary numbers
k	electro-mechanical coupling coefficient
L	inductance
M	figure of merit
N_c	frequency constant of a transducer
Q	quality ("pointedness" of a resonance peak)
Q_e, Q_m	electric and mechanical Q-factor
q, Q	electric charge
R	resistance
r	radius, radial coordinate
T	temperature
t	thickness
U	voltage
u	displacement
w	width
X	reactance
x	chemical mixing ratio, fraction
Y	admittance

y	spatial coordinate
Z	impedance
z	axial coordinate

List of Greek quantity symbols

Symbol	Description
Γ	motional capacitance constant
δ	normalised damping factor
$\epsilon_i^{\varepsilon}, \epsilon_i^{\sigma}$	relative permittivity at constant strain, at constant stress
ζ	damping ratio
η	loss factor
ν	Poisson's ratio
ϱ	density
σ_I	local stress intensity
ϕ, φ	phase angle
ω	angular frequency

List of abbreviations

Abbreviation	Description
AC	alternating current
APDL	Ansys Parametric Design Language
BC	boundary condition
BVD	Butterworth-Van Dyke
CFD	computational fluid dynamics
DOF	degree of freedom
EA	evolutionary algorithm
FE,FEM	finite element (method)
FSI	fluid-structure interaction
FV,FVM	finite volume (method)
IKET	Institute for Nuclear and Energy Technologies (Institut für Kern- und Energietechnik)
KIT	Karlsruhe Institute of Technology (Karlsruher Institut für Technologie)
ORNL	Oak Ridge National Laboratory
PZT	lead zirconate titanate (a piezoelectric ceramic)
RPI	Rensselaer Polytechnic Institute
RTV	room temperature-vulcanising (silicone)
SF	sonofusion

Appendix R

ANSYS APDL scripts

R.1 Modelling the piezoelectric ceramic with "plane223" elements

The piezoelectric transducer material was modelled with plane223 elements. These generic finite elements can handle various effects like piezoresistivity, electroelasticity and others through their coupled-field functionality. The setup of these elements in their piezoelectric mode was done following the example titled "VM-231" of the Ansys verification manual [9]. The corresponding APDL code was cast into the two macros listed below. That code results in a polarisation of the material parallel to the y-axis, i. e. the vertical axis of the drawing plane, so in the axis-symmetric case it corresponds to the z-axis of an r-z coordinate system. It can be noted that for the polarisation axis in the model to change, nothing in the two macros actually has to be modified because the polarisation axis change can be achieved by meshing the geometry while a rotated local coordinate system is activated (see section R.1.3), an approach allowing a simpler code structure.

R.1.1 A macro for material constants of PZT

```
pztflag=arg1
*if,pztflag,eq,'pzt4',then
  !piezo ceramic (PZT-4)
  *set,piez_dens,7500
  *set,piez_damp,-1 ! value of -1 serves as a flag not to use mp,damp
  *set,piez_dmpr,0.001 ! value of -1 serves as a flag not to use mp,dmpr
  *set,piez_lsst,0.04 ! value of -1 serves as a flag not to use mp,lsst
  *set,piez_perx,1475
  *set,piez_pery,1300
  *set,piez_perz,1475
  *set,piez_c11,13.9e10
  *set,piez_c12,7.78e10
  *set,piez_c13,7.43e10
  *set,piez_c33,11.5e10
  *set,piez_c44,2.56e10
  *set,piez_c66,3.06e10
  *set,piez_e15,12.7
  *set,piez_e31,-5.2
  *set,piez_e33,15.1
  *set,piez_d15,496e-12    ! PZT-4 according to O.B. Wilson and Landolt-Boernstein
  *set,piez_d31,-123e-12   ! PZT-4 according to O.B. Wilson and Landolt-Boernstein
  *set,piez_d33,289e-12    ! PZT-4 according to O.B. Wilson and Landolt-Boernstein
*elseif,pztflag,eq,'pzt8',then
  !piezo ceramic (PZT-8)
  *set,piez_dens,7550
  *set,piez_damp,-1 ! value of -1 serves as a flag not to use mp,damp
  *set,piez_dmpr,0.001 ! value of -1 serves as a flag not to use mp,dmpr
  *set,piez_lsst,0.01 ! value of -1 serves as a flag not to use mp,lsst
  *set,piez_perx,1290
```

```
  *set,piez_pery,1000
  *set,piez_perz,1290
  *set,piez_c11,13.7e10
  *set,piez_c12,6.97e10
  *set,piez_c13,7.16e10
  *set,piez_c33,12.35e10
  *set,piez_c44,3.14e10
  *set,piez_c66,3.37e10
  *set,piez_e15,10.4
  *set,piez_e31,-4.0
  *set,piez_e33,13.2
  *set,piez_d15,390e-12    ! Channel Industries C5800
  *set,piez_d31,-107e-12   ! Channel Industries C5800
  *set,piez_d33,245e-12    ! Channel Industries C5800
*endif
```

R.1.2 A macro for setting up "plane223" elements

```
axisymflag=arg1

!***************************
!***  define elements  *****
!***************************

*if,axisymflag,eq,0,then
  et,1,plane223,1001,0,0,0       !2-D element piezoceramic
*else
  et,1,plane223,1001,0,1,0       !2-D element piezoceramic
*endif

!***************************
!*DEFINE MATERIAL PROPERTIES*
!***************************

!setup of PZT-8 polarised in the positive y-direction
!This code is based on VM 231 of the verification manual.

*if,piez_dmpr,ne,-1,then
  MP,DMPR,1,piez_dmpr
*endif
*if,piez_damp,ne,-1,then
  MP,DAMP,1,piez_damp
*endif
*if,piez_lsst,ne,-1,then
  MP,LSST,1,piez_lsst
*endif
MP,DENS,1,piez_dens
MP,PERX,1,piez_perx ! PERMITTIVITY AT CONSTANT STRAIN
MP,PERY,1,piez_pery
MP,PERZ,1,piez_perz
TB,ANEL,1 ! ANISOTROPIC ELASTIC STIFFNESS
TBDA,1,piez_c11,piez_c13,piez_c12 ! c11,c13,c12
TBDA,7,piez_c33,piez_c13 ! c33,c13
TBDA,12,piez_c11 ! c11
TBDA,16,piez_c44 ! c44
TBDA,19,piez_c44 ! c44
TBDA,21,piez_c66 ! c66
TB,PIEZ,1 ! PIEZOELECTRIC STRESS COEFFICIENTS
TBDA,2,piez_e31 ! e31
TBDA,5,piez_e33 ! e33
TBDA,8,piez_e31 ! e31
TBDA,10,piez_e15 ! e15
TBDA,15,piez_e15 ! e15
```

R.1.3 Rotated polarisation via "aatt" command

A useful set of local coordinate systems can look like this:

```
local,11 ! piezo ceramic polarised in the y-direction (unchanged directions)
local,12,,,,,-90 ! piezo ceramic polarised in the x-direction
local,13,,,,,,90 ! piezo ceramic polarised in the z-direction
```

Then the rest can be accomplished by using the material attribution command `aatt` before meshing.

```
asel,s,loc,y,0,h ! select above origin
aatt,1,,1,11 ! later mesh these areas with material number 1 and in local coordinate system 11
asel,s,loc,y,-h,0 ! select below origin
aatt,1,,1,12 ! later mesh these areas with material number 1 and in local coordinate system 12
asel,allsel ! select everything
amesh,allsel ! meshing
```

R.1.4 Fluid-structure interaction via the "fsi" command

```
! assumption:   material type numbers are set in the materials macro with these variable names:
! m_piez=1 $ m_glass=2 $ m_epoxy=3 $ m_sili=4 $ m_liq=flet $ m_alu=9 $ m_steel=10 $ m_fixat=11
! acetone with only pressure DOF is number 5
! acetone with additional displacement DOF is number 6
! task: areas meshed with element type 5 need a layer of element type 6 where in contact with solids

! non-modified liquid in contact with glass
esel,s,mat,,5 $ nsle,s $ esel,all $ esel,s,mat,,2 $ nsle,r $ esel,all $ esel,s,mat,,5 $ esln,r
emodif,all,type,6 $ emodif,all,mat,6 $ sf,all,fsi $ allsel,all

! non-modified liquid in contact with silicone
esel,s,mat,,5 $ nsle,s $ esel,all $ esel,s,mat,,4 $ nsle,r $ esel,all $ esel,s,mat,,5 $ esln,r
emodif,all,type,6 $ emodif,all,mat,6 $ sf,all,fsi $ allsel,all

! non-modified liquid in contact with steel
esel,s,mat,,5 $ nsle,s $ esel,all $ esel,s,mat,,10 $ nsle,r $ esel,all $ esel,s,mat,,5 $ esln,r
emodif,all,type,6 $ emodif,all,mat,6 $ sf,all,fsi $ allsel,all

! non-modified liquid in contact with aluminium
esel,s,mat,,5 $ nsle,s $ esel,all $ esel,s,mat,,9 $ nsle,r $ esel,all $ esel,s,mat,,5 $ esln,r
emodif,all,type,6 $ emodif,all,mat,6 $ sf,all,fsi $ allsel,all

! non-modified liquid in contact with epoxy
esel,s,mat,,5 $ nsle,s $ esel,all $ esel,s,mat,,3 $ nsle,r $ esel,all $ esel,s,mat,,5 $ esln,r
emodif,all,type,6 $ emodif,all,mat,6 $ sf,all,fsi $ allsel,all


! already modified liquid in contact with glass
esel,s,mat,,6 $ nsle,s $ esel,all $ esel,s,mat,,2 $ nsle,r $ esel,all $ esel,s,mat,,6 $ esln,r
sf,all,fsi $ allsel,all

! already modified liquid in contact with silicone
esel,s,mat,,6 $ nsle,s $ esel,all $ esel,s,mat,,4 $ nsle,r $ esel,all $ esel,s,mat,,6 $ esln,r
sf,all,fsi $ allsel,all

! already modified liquid in contact with steel
esel,s,mat,,6 $ nsle,s $ esel,all $ esel,s,mat,,10 $ nsle,r $ esel,all $ esel,s,mat,,6 $ esln,r
sf,all,fsi $ allsel,all

! already modified liquid in contact with aluminium
esel,s,mat,,6 $ nsle,s $ esel,all $ esel,s,mat,,9 $ nsle,r $ esel,all $ esel,s,mat,,6 $ esln,r
sf,all,fsi $ allsel,all

! already modified liquid in contact with epoxy
esel,s,mat,,6 $ nsle,s $ esel,all $ esel,s,mat,,3 $ nsle,r $ esel,all $ esel,s,mat,,6 $ esln,r
sf,all,fsi $ allsel,all
```

R.1.5 APDL commands for solving the model

```
/solu
antype,harmic
seltol,1e-6
nsel,s,loc,x,or+epd $ nsel,r,loc,y,piyd,piyu $ d,all,volt,0 $ nsel,all       ! grounding inner electrode
nsel,s,loc,x,or+epd+pid $ nsel,r,loc,y,piyd,piyu $ d,all,volt,u $ nsel,all   ! loading outer electrode with voltage u = 100
harfrq,fini,fend    ! frequency range
nsubst,nsteps  ! number of frequency steps
kbc,1 ! stepped loads
eqsl,front
solve
finish
```

List of abbreviations

Abbreviation	Description
APDL	Ansys Parametric Design Language
FSI	fluid-structure interaction
PZT	lead zirconate titanate (a piezoelectric ceramic)

Appendix S

More FEM simulation results

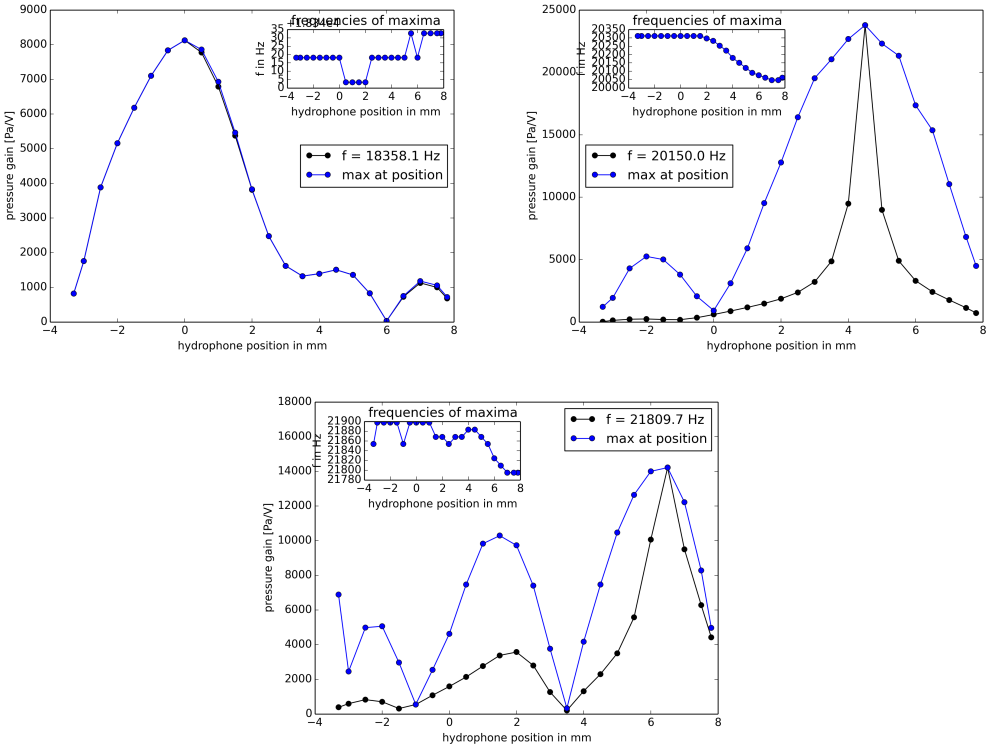


Figure S.1 Sound pressure profiles of the FE model of resonator Nº 5
These profiles show the sound pressure amplitude mode shapes of the three resonances exhibited by the FE model of the opened resonator 5 in the frequency interval of interest. The underlying dataset is plotted as a colour map in figure Q.7, from where it can be seen that the hydrophone position influences the resonance frequencies. This also means that mode shape plots cannot be created by simply reading the pressure data from the FE mesh nodes along the central axis from one single simulation. Instead they have to be compiled from many simulations with varied hydrophone positions gathering the pressure loaded onto the surface of the hydrophone tip. These mode shapes have to be compared to their experimental counterparts, the profile plots based on the hydrophone data, which can be found in figure O.24 on page 387 and P.22 on page 414.

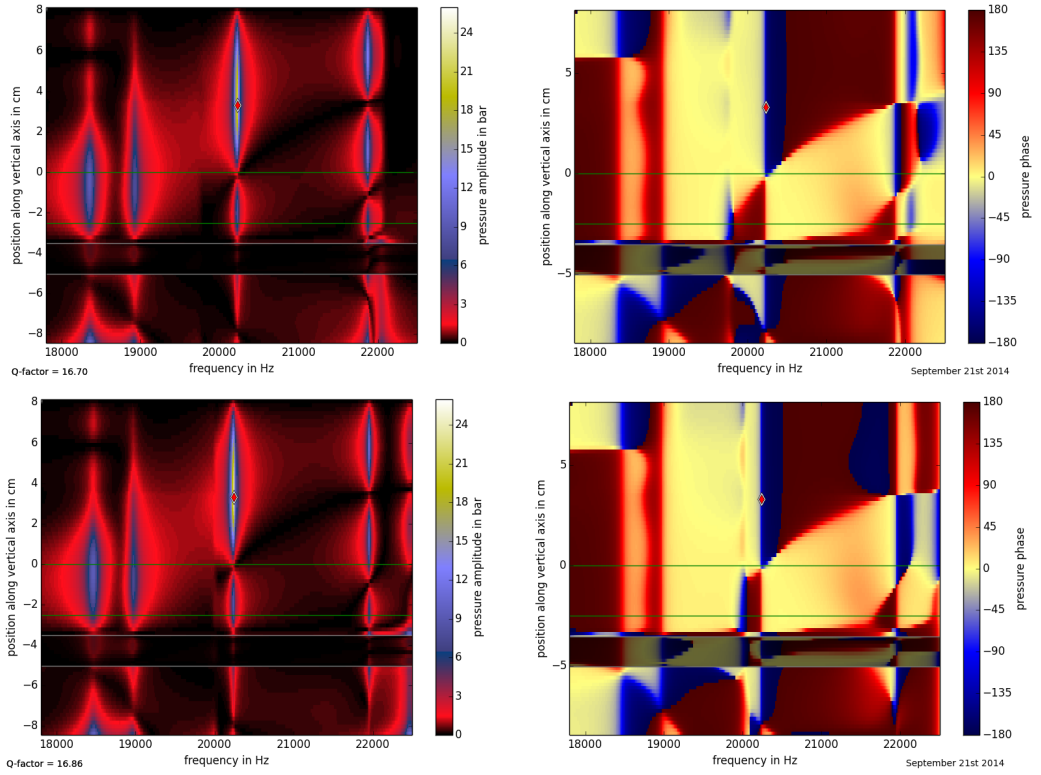


Figure S.2 Switching between the two epoxies
Two-component epoxy is used for glueing the transducer to the glass wall. Two sets of material constants [69, 497] were tried in the FEM simulations, they are given in table Q.3. The above sound pressure maps show that there is a visible impact when switching between the two sets with the most substantial change around the 3rd resonance. The upper maps show the case with the Young's modulus of $1 \times 10^9\,\mathrm{N/m^2}$ which was kept as the default value for all other FEM simulations, whereas the lower maps show the landscapes yielded by a Young's modulus raised to $5.86 \times 10^9\,\mathrm{N/m^2}$. This shifts a weak structure from 19.8 to 20 kHz and a side-peak of the 3rd resonance from 22 to 22.3 kHz. All other settings are equivalent to the simulation presented in figure Q.7 for the comparison with the real-world counterpart of resonator 5. As the two structures being shifted are completely absent from the hydrophone data plotted in fig. Q.7, there is no hint about which of the two epoxy alternatives is closer to reality.

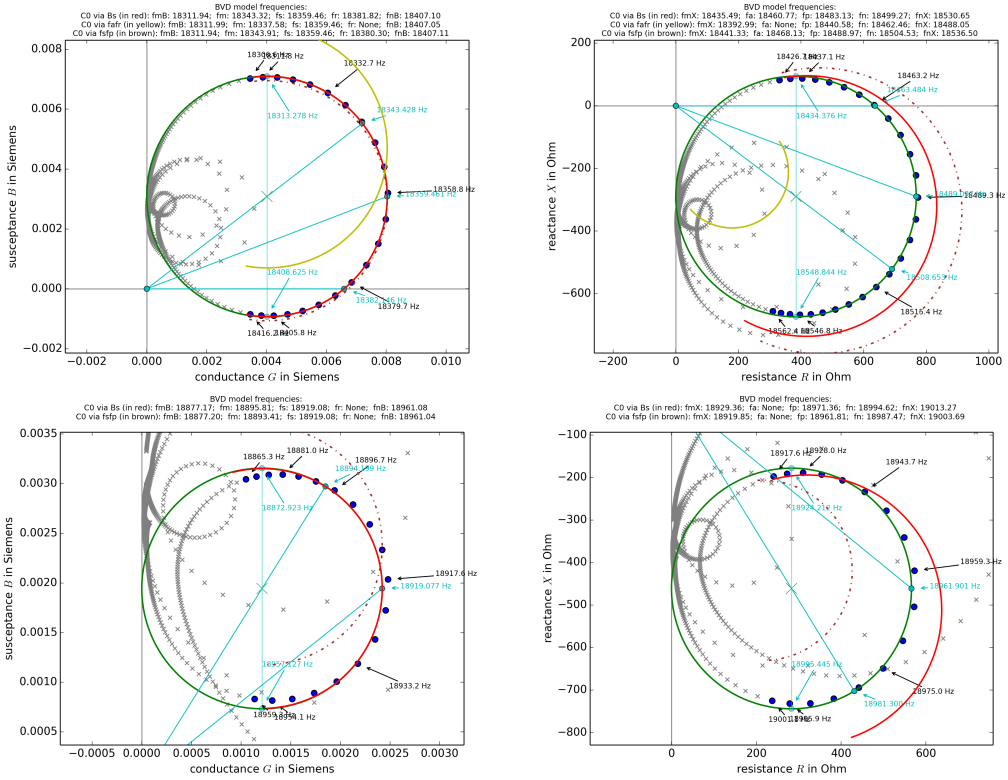


Figure S.3 Y- and Z-circles of the FE model of resonator № 5: 1st resonance
Fitted Y-circles (left column) and Z-circles (right column) based on the simulation results gained with the FE model of the opened resonator 5 are shown for the two peaks of the split 1st resonance. The simulation settings are the ones of the 9th row in the pressure map of figure Q.7 where the hydrophone position is 3.5 cm. The blue dots are the data points used for generating the fitted circles shown in green. The criterion for selecting the blue data is $G > 0.4G_{\max}$ of the corresponding resonance. The grey crosses show the entire dataset of the frequency sweep from 17.8 to 22.5 kHz. The BVD circuit model quantities deduced from the fitted circles are given in table Q.5, while the BVD model Y- and Z-circles are directly plotted over the data. Since the plots indicate the offset of the BVD model circuits in the Y and Z plane but not those on the frequency axis, the characteristic frequencies of the BVD models are given at the top of each plot. There are at most three BVD models for each resonance because the parallel capacitance C_0 can be either calculated as $C_0 = B_s/\omega_s$ (red circles), $C_0 = f_r^2 C/(f_a^2 - f_r^2)$ (yellow), or $C_0 = f_s^2 C/(f_p^2 - f_s^2)$ (dash-dotted brown).

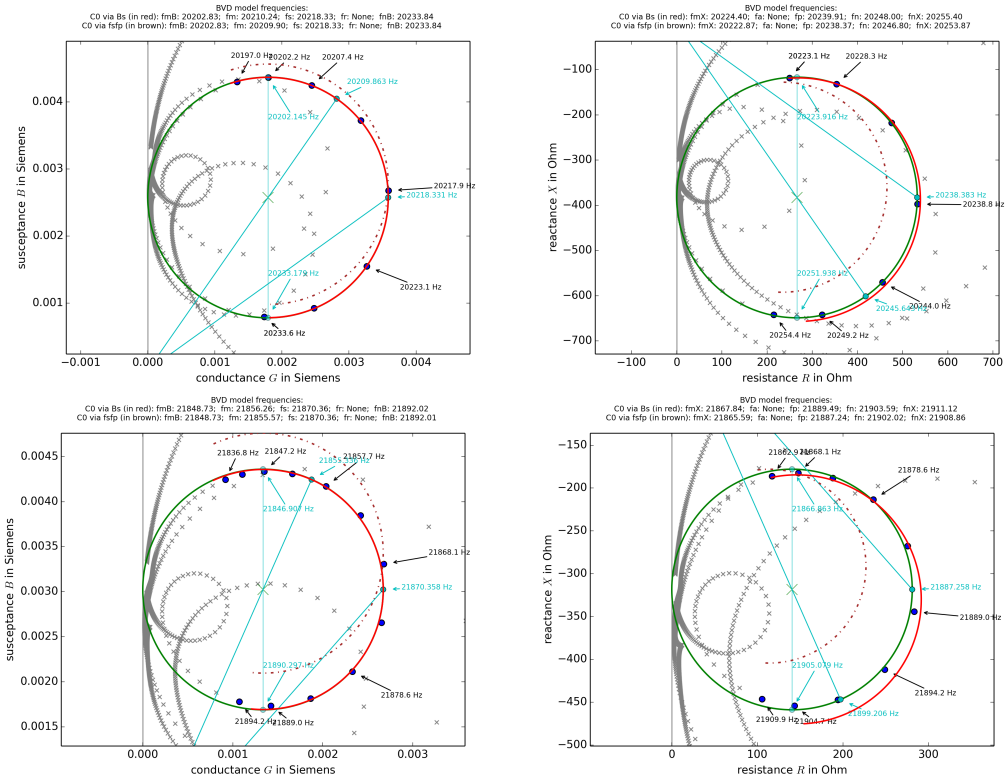


Figure S.4 Y- and Z-circles of the FE model of resonator № 5: 2nd and 3rd resonance

As for these two resonances the circles do not intersect with the real axis, there are no characteristic frequencies f_r and f_a. Therefore, the yellow BVD models based on values of C_0 gained via the formula $C_0 = f_r^2 C/(f_a^2 - f_r^2)$ are missing here. For the selection of the data to be fitted (blue subset) the threshold of $G > 0.35 G_{\max}$ has been used here.

Lists of symbols and abbreviations

List of symbols

Symbol	Description
B	susceptance
C	capacity
f	frequency
G	conductance
R	resistance
X	reactance
Y	admittance
Z	impedance
ω	angular frequency

List of abbreviations

Abbreviation	Description
BVD	Butterworth-Van Dyke
FE,FEM	finite element (method)

Appendix T

Global optimisation with evolutionary algorithms

Mother Nature's evolutionary algorithm is responsible for some amazing solutions to engineering problems like spider webs, shark skin, bird wings with feathers, rotating motors of bacteria flagellas, eyes, or brains. Nature solves these problems by composing source code written into DNA and RNA that gets translated into proteins. The proteins put each other into place and some of them produce and process other molecules, so that in the process complex systems like bacteria, spiders, trees, warthogs, and humans get built. After computers had widely become available, programmers developed interest in replicating with lists of zeros and ones what nature is doing with the four-letter alphabet of nucleic acids. Sometimes the focus was on understanding how nature works by simulating simplified versions of it, sometimes it was on using the lessons learnt from the theory of evolution for devising algorithms capable of solving problems that can be expressed as computer programs. More than five decades of research in evolutionary computation (EC) have led to a vast array of evolutionary algorithms (EA) where the most popular and well-studied branches are evolution strategies (ES), evolutionary programming (EP), genetic algorithms (GA), simulated annealing (SA), genetic programming (GP), memetic algorithms (MA), ES with covariance matrix adaptation (CMA-ES), differential evolution (DE), and swarm algorithms like ant colony optimisation (ACO) or particle swarm optimisation (PSO). Simulated annealing and the swarm algorithms show that besides the theory of evolution also other fields like thermodynamics or swarm behaviour lent inspiration to the field. EAs as function optimisers can be classified either under stochastic optimisers or metaheuristics (MH). The term evolutionary computation (EC) is somewhat broader and also encompasses concepts like neural networks (NN) and learning classifier systems (LCS). To be in EC, something must have incorporated the principle of *survival of the fittest* [299]. To be an EA, an algorithm needs to act on and evolve symbolic representations (*genotype*) of problem solutions (*phenotype*). To be able to speak of a metaheuristic, there must be a distinction between an algorithm with overarching control forming an upper level and a toolbox of simple operations or subroutines (i. e. a portfolio of heuristics) on the lower level. Describing an EA as a metaheuristic implies the notion that the EA is conceptually independent from any particular choice or implementation of the low-level routines.

This appendix chapter provides the background information to chapters 3, 4, & 5 where it is described *how* two evolutionary algorithms (EA) were used for resonator optimisation. The background chapter should describe *what* EAs are, including *where* they have to be placed among the many existing classes of optimisers, and *why* they work efficiently and have been favoured over other types of search and optimisation algorithms. The *what* and *why* of the facets of EAs are of course interdependent making it hard to treat them one after the other. It has been tried to solve the dilemma by proceeding in three steps, first by implicitly describing the *why*, and a bit of the *what* at a fundamental level by looking at solving engineering problems in general and pointing out its only sometimes disrupted inherent iterative and evolutionary character, next by describing *what* the fundamental mechanisms of biologic evolution are and *why* they drive the process, and finally by describing a selection of classic and modern EAs, still paying attention to how the ingredients (*what*) reflect the purposes (*why*).

T.1 Introductory remarks on numerical global optimisation and EA

T.1.1 Solving engineering problems

The core of each solution to an engineering problem is a good idea. Depending on the type of problem the idea may be just the general shape of a structural component, for example the shape of a nail, but at other times it can be a concept of how to combine subsystems into a complex system, like the idea of using a turbine instead of a piston engine to make a helicopter more efficient. Looking at how even such a simple thing as a nail has changed, or better yet *evolved*, since its invention in prehistoric times, and at how many shapes of it are in use today, shows that a process of problem-specific fine-tuning belongs to each optimal-solution search.

Inventions start with imagination and analytic thinking employing all remembered experiences and rules (exact, induced, or thumb rules) in order to come to an idea of improvement. This is sometimes also called *heuristic*[1] *reasoning*. Then follows the implementation. For tiny engineering problems or repair work, that's it. Most other engineering tasks are ones that have to be solved over and over again. Think of hand axes and boats as two more examples. New exemplars have to be made when the old ones break or become too inefficient. For the Stone Age craftsman, curiosity, new sorts of stones found or traded, growing experience, modified or new modes of using the tool, all this will lead him to try out various forms of the hand axe throughout his life, and to establish an ever more sophisticated and comfortable sequence of production steps. The example is to show that innovative engineering consists of disruptive ideas (the first hand axe, railway, jet engine ever) and less disruptive variations and modifications of ideas or techniques. The repetitive cycle

[1] From the Greek εὑρίσκειν, 'to discover, find'. Heuristic, according to Polya [358], is an almost forgotten "branch of study [...] belonging to logic, or to philosophy, or to psychology" with the aim "to study the methods and rules of discovery and invention", and it can be rephrased as *ars inveniendi* (lat.), the art of inventing. It contains two notions, the one of free imaginative thinking and the one of methodology.

responsible for the evolution steps of one invention could be broken down to the loop shown in figure T.1 on the left which consists of heuristic thinking leading to a concept, followed by implementation and usage (i. e. experiments) and judgements, whereafter the cycle can rebegin. As a result, the various shapes of prehistoric hand axes often can tell archaeologists about the time frame of their production and the experience and proficiency of the craftsman who produced it. The other example, boats, can illustrate how evolution steps differ in being more disruptive (invention of new types of sails, introduction of propellers) or less (consecutive small variations of hull shapes). But it can also show that good imagination and ideas are not everything, that modernity brought the necessity of a different kind of work: at some point scholars were needed who knew how to model physics and could calculate whether a new big ship would be instable and in constant danger of tipping over or not, and secondly, who knew numerical recipes to deal with parameter variations (e. g. iterative schemes like Newton's method) to settle trade-offs between conflicting goals like making the boat stable *and* reducing its depth of flotation. Looking at the right hand side of figure T.1, applying an iterative numerical recipe for finding an optimal solution of such a trade-off would correspond to the small optimisation loop (b). Checking the calculations with a small model boat and making further modifications would be the optimisation cycle (a). Judging experiences made with the ready-built boat on the ocean will then close the large innovation cycle. Today, computer-based numerical simulation and optimisation methods are two branches complementing experimentation and mathematical analysis. They have substantially gained in importance over the past decades. Numerically calculating ballistic trajectories enabled the leap into space half a century ago, and today, not many components will be found in a power plant, aeroplane, or car, that have not been subject to some sort of simulation or optimisation.

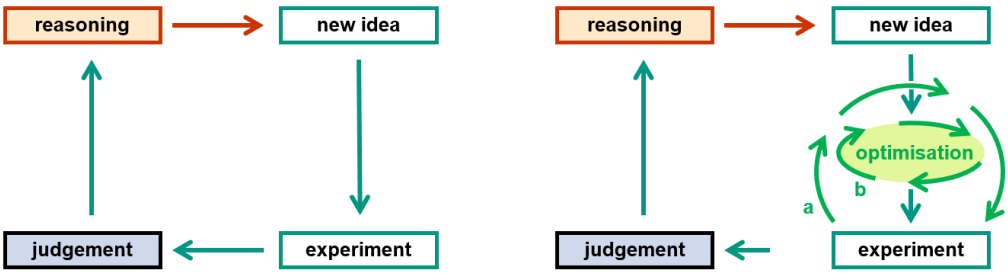


Figure T.1 Why solutions evolve.
Most engineering solutions are needed not just once. A blacksmith will forge a door hinge not just once, same with a hunter and a bow, and a carpenter will learn dozens of ways to connect two wood beams and implement thousands of connections in a life. The cycle, shown on the left, of creative thinking, trial, error, critical thinking, and renewed trials is the basis on which engineering solutions evolve (and our civilisation). In larger projects, a small part of the solution may be picked out and tested in a separate environment. For example, several versions of a bolt might be examined on a test stand before only one of the versions ends up being used in the large system (e. g. an aeroplane). This would correspond to the development cycle depicted on the right with a subordinate experimental trial-and-error scheme, the optimisation cycle (a), for the bolt. Staying with the example of aeroplane development, the optimisation cycle (b) could stand for a purely conceptual optimisation cycle like discussing several body shapes and engine placement versions, but it could also stand for the simulation-aided optimisation of subsystems or facets of the whole system, like the simulation-aided aerodynamic optimisation of the hull. The outer cycle would in that case be the aeroplane model development cycle. From today's perspective, it looks nearly impossible to develop a complex system like an aeroplane without many such subordinate optimisation cycles targeted at improving components, subsystems, and subsystem interactions.

Numerical optimisation means an algorithm is automatically deciding (repeatedly) which of two competing solutions is better. This makes it hard to apply numerical optimisation to a hand axe or a saxophone, because how well such things are received by a human user depends on the combination of a large amount of physical, haptic, and aesthetical information and is practically impossible to model on a computer. But with other examples it looks different. Automatic parameter tuning is promising when there are clear physical criteria forming the main goal, like the energy conversion efficiency of water and wind turbines, the weight of a bicycle frame, the efficiency of an antenna. The two keys of solving an optimisation problem numerically are to come up with a useful objective function, or a small collection thereof, and to choose or build a suitable algorithm searching the landscape unfolded by the objective function in an efficient manner. Let us start with a proper definition of the parameter optimisation problem.

T.1.2 Formal definition of the parameter optimisation problem

The real-parameter optimisation problem

A real-parameter optimisation problem is the search for an optimal combination of n input parameters $(x_1, x_2, ..., x_i, ..., x_n) = \vec{x}$ resulting in the best achievable value of one or more quantifiable criteria $f_{\text{obj},k}(\vec{x}) : \mathbb{R}^n \to \mathbb{R}$. Depending on the nature of a physical criterion (weight, stability, power, cost), either maximisation or minimisation can be desired, but without loss of generality[2] only minimisation problems need to be considered in the mathematical formulation. Furthermore, throughout most of the chapter only single-objective problems will be discussed and the index k can be dropped. Labelling the set of all possible input vectors of the function f_{obj} with $\mathcal{X}$ and the ensemble of all reachable output values with $\mathcal{Y}$, the projection can be symbolised with

$$f_{\text{obj}} : \mathcal{X} \to \mathcal{Y}, \tag{T.1}$$

where $\mathcal{X}$ is called the domain of f_{obj} and $\mathcal{Y}$ its co-domain. Generally, an optimisation problem can be constrained and solutions violating one or more constraints are considered inacceptable independent of their objective function value (e. g. when a design breaches some imposed temperature, material stress, or cost limit). Considering that any constraint inequality expression $g(\vec{x}) < b$ with a threshold value $b \in \mathbb{R}$ can be re-arranged in the form $g'(\vec{x}) < 0$, the task of constrained optimisation can be described as

$$\begin{aligned} &\text{minimise} && f_{\text{obj}}(\vec{x}) && \\ &\text{subject to} && g_l(\vec{x}) < 0, && l = 1, ..., m \end{aligned} \tag{T.2}$$

where $f_{\text{obj}}(\vec{x}) : \mathbb{R}^n \to \mathbb{R}$ is the objective function and the functions $g_l(\vec{x}) : \mathbb{R}^n \to \mathbb{R}$ are the constraint functions. Sometimes optimisation problems are defined so that the domain $\mathcal{X} \subset \mathbb{R}^n$ encompasses only those input vectors where the parameters x_i fulfil all constraints expressed among them. One can imagine for instance geometries

[2] from now on sometimes abbreviated as "w. l. o. g."

becoming infeasible if parts overlap and a distance becomes negative, conditions which can render the evaluation of the objective function in principle impossible. Any reduction of the search space by pure "input constraints" may be helpful, in particular if the geometric shape of the resulting domain $\mathcal{X}$ can be understood and allows gearing an algorithm accordingly. But what if constraints involve output quantities becoming available only after evaluation of a solution candidate $\vec{x}$? In real-world optimisation tasks it is often the case that a function $g(\vec{x})$ is not a simple algebraic function of the parameters x_i, but rather represents measurements like temperatures or stress levels as stability, safety, or failure criteria probed in the simulation or lab setup. Handling nontrivial constraints in EAs is an own issue which has been avoided in the present work because a simple work-around consists in representing constraint violations as additive penalty terms in the objective function. Therefore, only the subclass of *unconstrained* minimisation problems that can be formulated as

$$\begin{aligned} \text{minimise} \quad & f_{\text{obj}}(\vec{x}) \\ \text{with} \quad & x_i \in [a_i, b_i], \qquad i = 1, ..., n \end{aligned} \tag{T.3}$$

will be treated here, where the only restriction of the search domain $\mathcal{X} \subset \mathbb{R}^n$ comes in the form of lower and upper boundaries along each dimension. Due to the easy feasibility of projections such a domain definition can w. l. o. g. be considered equivalent to $\mathbb{R}^n$ or a unit cube $[0, 1]^n$. It can be useful to keep in mind that an algorithm designer may jump between domain representations as suitable and convenient for algorithm implementation. Let us call the sought location of the global optimum $\vec{x}_{\text{opt}}$ and the extremal function value f_{opt}.

Evolutionary algorithms are often intended for use in minimisation problems when there is no gradient information available (i. e. ∇f_{obj} is not defined or impossible or too costly to calculate) and generally for cases where knowledge about the morphology of $f_{\text{obj}}(\vec{x})$ is low[3]. This is in accord with the trial-and-error nature of engineering problems and the fact that in simulation-based optimisation a computation of the gradient cannot always easily be built into a given routine for the computation of f_{obj}. Therefore, explicitly, no further requirements on $f_{\text{obj}}(\vec{x})$ are formulated, it can contain sharp ridgelines, cliffs, noise, singularities, anything, because for the optimising algorithm the engineering problem to be optimised is regarded as a black box generating an output value f_{obj} from the input parameters x_i, of which it should not matter whether they stand for distances and angles in a structural design, or whether they characterise the specs of resistors, capacitors, coils, diodes, etc. on a circuit board. The *black box* or *blind search* character of the search problem can be illustrated very well by figure T.2.

As will be explained below, the search for the global minimum can become quite tedious in higher dimensions, so tedious, that finding it is of no practical relevance. Then, finding lower values more quickly than competing procedures is the only

[3]Note the use of the word "low" instead of "nonexistent". The rule that small variations of $\vec{x}$ ensue small changes in f_{obj} holds most of the time in engineering problems as well as in biologic evolution. This is already a substantial piece of information. The question whether EAs are a good choice for any arbitrary problem with completely unknown f_{obj} is embedded in a wide range of literature on *no free lunch theorems*, see below in section T.1.4.

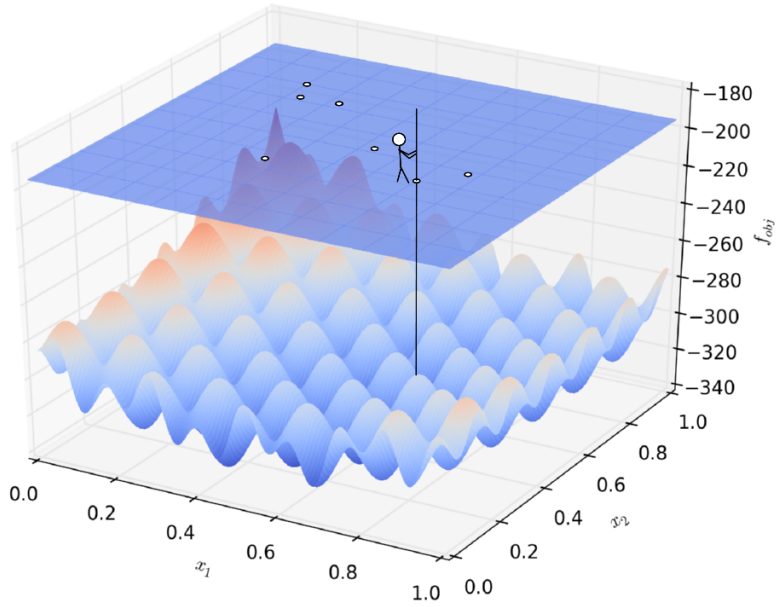


Figure T.2 Black-box search without derivatives.
This illustration (based on the one used by I. Rechenberg in his lectures on evolution strategies [380]) describes the blind search character of the derivative-free function minimisation or maximisation problem. Those search problems can be called easy, where it is possible to infer valid assumptions on the overall look of the whole landscape with just a few samples and where these assumptions are valuable and helpful for finishing the search.

thing that counts for a function optimiser. Maybe one should distinguish between *melioration*[4] or *amelioration* and *optimisation*, where the former means seeking ever better qualities and the latter the mathematically stringent search for the global optimum (see e. g. Beyer et al. [40, 42, 43]). Conventionally, however, the term optimisation is used for both cases.

Remarks on discrete variable optimisation

A discrete-variable optimisation problem exists when the parameters x_i are restricted to take values from discrete sets like $[0, 1]$, $[-1, 0, 1]$, $[1, 2, 3, 4, 5, 6]$, $\mathbb{N}$, a binary code like $[000, 001, 010, 011, ..., 111]$, a hexadecimal code or similar. Culberson[5] [98] gives a perfect example of a discrete blind search problem: two persons are sitting on a park bench, one of them being blind. The blind one is holding a Rubik's cube, gives it a twist, holds it up to his friend, and receives the answer "No." The blind person goes on by twisting the cube again and again, just to receive more and more negative answers.

There are other examples of discrete problems often used in academia. One of the simplest is the **onemax** function, a classic test problem from the field of genetic algorithms (GA), where a four-dimensional search vector would look like $(1, 1, 0, 1)$

[4] from Latin *bonus*, *melior*, *optimus*: good, better, best
[5] referring to a scene in the movie "UHF", USA 1989, director: Jay Levey

or $(0, 1, 0, 1)$ and the goal is to maximise the sum, i. e. to find the vector $(1, 1, 1, 1)$. It is a perfect example of illustrating how an incredibly simple task turns into a computationally costly problem once it is being conducted in the *two-men-on-a-park-bench* manner. Another classic example is the **travelling salesman problem (TSP)** where the shortest route is to be found for visiting n cities once and return home. In the **knapsack problem** a backpack has to be packed efficiently. In the simplest case a rectangular volume has to be filled with a subset of available rectangular pieces each attributed with a mass and a price tag such that the total value of the load is maximised without breaching a total weight limit. In **graph colouring** one has to imagine a map of countries and seek a colouring pattern so that across each borderline the colour changes, but the overarching task is to find out how many colours are at least needed. In general, graph theory and boolean logic are rich hosts of **combinatorial optimisation problems**. Then there are problems of **resource allocation and scheduling** like **vehicle routing (VRP)** or **job shop scheduling (JSSP)**. An example of the first is the weekly supply of m supermarkets with n trucks leaving from a central depot, an example of the latter is distributing m jobs of varying size onto n identical machines and minimising the total time.

Note that there is a big difference in just discretising a search space (e. g. tuning n distances between 10 and 80 cm in steps of 1 cm, or alternatively that range can be divided into 63 steps and $\vec{x}$ can be given as a list of six-digit binary numbers) where for any component of $\vec{x}$ any value is allowable and *combinatorial optimisation* like e. g. TSP where only permutations of one sequence are allowed or other constraints relating the vector components must be fulfilled. In the search space of the first example one can move sideways in any direction to get to a new allowable point. In the TSP one solution will be a list of cities like (d, a, c, b) and a neighbouring point in the search space is said to be a list that can be generated by swapping two cities. In this latter case neighbourhood distances in $\mathcal{X}$ are not euclidean distances but the number of city swaps needed to get from one list to the other. So one could say combinatorial problems are those discrete searches where there are additional constraints relating the search vector components. Or, the other way round, that those discrete problems, where the search space allows the interpretation as a euclidean space, are one special subset of the larger class of combinatorial problems (the latter view is reflected in [512]). The TSP can also make it clear that the issue of solution representation and problem-specific operations on solution candidates must always be of central interest when discussing search algorithms. A last general thought that can be taken from the TSP is on the topology of $f_{\text{obj}}(\vec{x})$ in combinatorial problems. A city swap has a smaller impact when the cities are close neighbours than when they are far apart. Therefore, in the normal case when there are more far than close neighbours to one city, one can expect that many neighbouring points in $\mathcal{X}$ will differ greatly in their objective function values. This is a big difference to many real-world continuous-domain problems where it can be assumed that most of the time small variations in $\vec{x}$ will lead to only small changes in $f_{\text{obj}}(\vec{x})$, where thus $f_{\text{obj}}(\vec{x})$ can often be assumed to be piecewise smooth. This is important for the discussion of the *no free lunch theorems* in section T.1.4 below.

Discrete parameter-tuning problems (not combinatorial problems) could in prin-

ciple be declared to be a special subclass of real-domain problems where f_{obj} contains an additional step of rounding each x_i to an integer, i. e.

$$\begin{aligned} & f_{\text{obj}} = g \circ h \\ \text{with} \quad & h : \mathbb{R}^n \to \mathbb{Z}^n \\ \text{and} \quad & g : \mathbb{Z}^n \to \mathbb{R}. \end{aligned} \tag{T.4}$$

Thus having discrete problems dissolve in real-parameter optimisation can be taken as the inverse of the view of binary-coded genetic algorithms (BCGA), where real-domain optimisation is seen as a limit case of increased-resolution binary parameter coding.[6] This can sometimes be helpful to separate the conceptual level of the search algorithm from the level of parameter representation, but blurring the distinction between continuous and discrete optimisation that way can also lead to those kinds of misconceptions that make the discussion of the impact of *no free lunch theorems* so controversial.

T.1.3 Objective functions in engineering problems

Finding or composing objective functions

Optimisation in engineering always has something to do with balancing competing goals, like weight versus stability or cost versus efficiency. The left graph in figure T.3 shows a generic one-dimensional trade-off situation: there is one design parameter x and two goal functions, criteria A & B, to be minimised, each one tearing the design parameter in a different direction. An example could be determining the engine size for an aeroplane of given dimensions. Here, the minimisation of the weight is competing with the maximisation of the thrust. The climbing rate can be taken as a scalar quality measure inherent to the physical system. A too small engine will not be able to pull up the plane, and with a huge engine and too small wings and propeller the plane will not lift either. Thus, the maximal climbing rate can be multiplied with -1 and taken as the objective function, making the application of a generic function minimiser possible. A very similar 1D optimisation example is the sizing of the catalyst, the exhaust cleaning facility, of a coal-fired power plant. The two competing goals are the cleaning efficiencies for SO_2 and NO_x versus the energy consumption of the catalyst, the drag on the energy efficiency of the plant. But here there is no inherent physical quantity telling about the optimality of the trade-off. A deliberate decision has to be made on the weighting of the two goals, and in the case a (single-objective) optimisation algorithm were to be applied, the decision would have to come in the shape of a mathematical formula projecting the two variables into one scalar objective function.

The right graph in figure T.3 describes another 1D problem with a primary goal to minimise (the black curve) and a threshold constraint. An example could be finding the optimal thickness of a steel rope: there is a given maximal force it has to bear, and the optimal rope thickness has to be found which yields the minimal material cost while ensuring that a critical level of material stress is not exceeded.

[6] Implementing a real-domain algorithm on a digital computer means in fact always the transition to a representative discrete-domain algorithm in a search space of finite resolution.

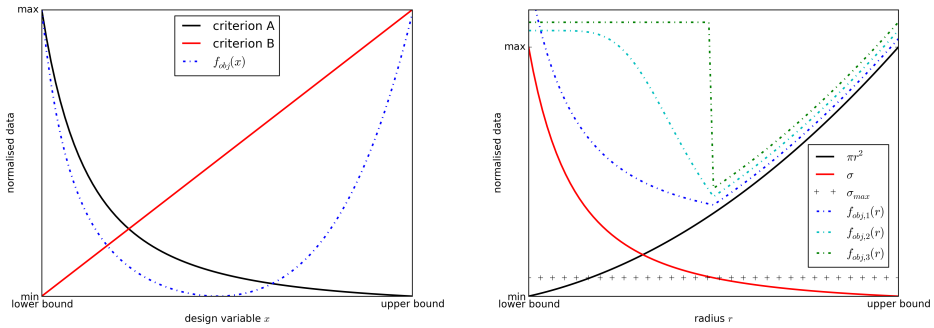


Figure T.3 Objective functions in engineering problems.
The plot on the left shows a generic trade-off situation with two criteria or subordinate goals, where the underlying system offers a meaningful physical quantity manifesting the overall solution quality, which can then be taken as the objective function. The plot on the right stands for a different situation, where an overall quality measure has to be engineered in a way to prevent an optimiser from finding solutions which are undesirable (where the curve σ is above the threshold marked by the crosses). Here, three possible objective function versions are plotted as dotted lines (with slight vertical offset for better visibility) which are all based on the black curve (criterion A) as the principal measure of solution quality to be minimised, but with deviations from that curve in order to penalise the undesired points.

The primary optimisation problem, i. e. minimising the cost of the rope of a given length, is trivial, it is solved by reducing the design variable, the rope radius r, to the lower bound. The secondary goal, i. e. keeping material stress below a critical value, is represented in the graph by the variable σ plotted in red and the threshold indicated by the crosses. The optimal choice is of course the smallest safely acceptable rope diameter and this is where the red line hits the threshold. If too thin rope diameters are not excluded by the choice of design parameter boundaries, then the prohibition of designs yielding exceeding stress levels must be somehow incorporated into the objective function. This can be accomplished by adding a penalty to the objective function. The plot shows actually three implementation versions of the penalty as dash-dotted lines. The first version adds a number proportional to the constraint transgression depth, the second version makes the objective function jump to a large default value wherever the threshold criterion is breached, and the third version is a steady transition to the large default value. It will later be seen that the time-development of a population-based EA in the search space can be seen as a swarm in motion expressing its own quasi-physical behaviour. If the penalty is understood as a tool to bounce the swarm back into the domain of desired solutions, it becomes obvious that its shape can have important effects.

Caution with objective functions creating unintended effects

Let us go next through the thought experiment of a shipyard engineer deciding to apply numerical optimisation to the century-old traditional hull shape of a boat. It can illustrate common traps of outsourcing design work transferring it from an intelligent human to a computation automat. We assume the engineer sets up a slim 2D calculation dealing with the geometry of the cross section of the boat, and how the masses of structure, cargo, engine, and displaced water play out and create a torque either stabilising the boat or tending to tip it over sideways. The design

parameters determining the boat's cross section geometry, being the input to the calculation, can now be optimised to make the boat more stable. But then the choices begin. The optimisation algorithm might for example be targeted at maximising the torque pushing the boat upright again when it has been tilted sideways by a given angle α_1, say, $\alpha_1 = 3°$. Alternatively, the engineer may decide to maximise the limit angle α_{lim} beyond which the boat capsizes, i. e. the angle up to which the stabilising torque turning the boat upright again is positive. Both are quantities of important physical meaning, offering simple and clear comparison measures between competing boat designs. Probably he will learn from looking at final results from both types of optimisations and come to a first conclusion, namely that a design has to be better in both criteria in order to be accepted by the algorithm for replacing inferior designs. The algorithm would have to be modified to be able to deal with multi-objective search. The Pareto-optimiser would have the task to reveal the topology of relevant segments of the Pareto front. In particular, the engineer would like to identify regions where one goal can be advanced substantially by sacrificing marginally on the other one. After that, another step could be to return to a single-objective function for the fine-tuning and for that purpose a combination of both criteria would have to be invented (e. g. by a product or weighted sum or by adding a penalty from transgressing a threshold in one criterion to the objective function based on the other one). After that, a typical user experience could then fall into the next trap by finding a boat shape yielding fantastic values for the two criteria, high turn-back torque at three degrees inclination and capsizal only beyond 70° inclination, but what if the magnitude of the stabilising torque is just an epsilon above zero between 5° and up to 70° from where it quickly turns to large destabilising values? In that case the optimisation algorithm would have succeeded in serving the word of the optimisation law, but not the spirit of the search for a safe boat. This hypothetical real-world numerical optimisation user story would only come to a good end after several learning experiences and after distilling a complex compound objective function reflecting the many important criteria: depth of flotation limited to common values, maximised cargo volume at a given breadth, stabilising torque not below an angle-depending critical value for all sideways tilting angles between zero and a relevant limit angle, and after all that, the boat's hull shape might still be disadvantageous in terms of drag, because this quality measure was not asked for in the optimised model and would require a wholly different model of the fluid dynamics around a 3D hull. Therefore, in a real-world shipyard, one will go through a cycle of both calculations, the one for capsizing stability and the one for drag with each new relevant design proposal iteration, and whether a step of numerical optimisation can usefully form a part of the process will depend heavily on the care taken with respect to the details of the implementation approach.

As the remaining chapter will be dedicated solely to the interplay of EAs and topologies of objective functions assumed as given and unchangeable, the just described example is to raise awareness that an equally important part of the problem solving can lie hidden in the choice of the objective function. The example shows that some raw and ad-hoc-chosen objective functions can feature valleys misleading the search algorithm into regions of the design space not intended by the user and that by a more thoughtful composition of $f_{\text{obj}}(\vec{x})$ the search landscape can be made

much less misleading. This should be borne in mind, because it means that before going at length to come up with an appropriate EA for a very hard optimisation problem, some thoughts should be spent on seeking possibilities to tweak $f_{\text{obj}}(\vec{x})$ in ways making the search easier and perhaps making the problem solvable at once by less costly and much simpler types of search algorithms.

T.1.4 What makes optimisation problems hard

Figure T.4 qualitatively shows various objective functions in one and two dimensions illustrating different difficulty levels for minimising algorithms. In the top row of 1D problems the plot on the left contains three landscapes with the same global minimum and no local minima, no saddles, so that the valley has monotonically ascending slopes. These preconditions make it possible to find the global minimum starting from any point in the search space and just following the gradient.

The plot in the centre shows the Rastrigin function in 1D,

$$f(x) = A + x^2 - A\cos(2\pi x) \quad \text{with } A = 10.$$

It is based on a parabolic function where a wavy structure has been introduced by adding a cosine signal. A gradient method will find the global optimum of this landscape only if the search starts on the slopes of the central valley. Nevertheless, the global minimum search of this function is not hard at all. If the features of the landscape are known, simple search strategies can be devised to exploit them, like following the envelope of the minima, the envelope of the maxima, or following the local average after filtering out the ripples with a low-pass. It is worth noting that in order to get useful information from the filtering step, the sample points do not have to lie densely in the search space. Probing the function on uniformly distributed random points will give reliable average gradient information if only enough samples are taken – completely independent of how many samples per ripple wavelength are taken.

The 1D function on the right shows a much more difficult search scenario, where there is no readily interpretable average slope information guiding towards the global minimum in the centre. The landscape is obscured by ripples of various wavelengths, the function is unsteady at $x = 0.2$, and noise dominates the part on the right.

The second row of plots shows colour maps of search landscapes in two dimensions. The left plot shows the Rastrigin function which is defined in n dimensions as the linear superposition of n 1D functions along the n coordinate axes.

$$f(\vec{x}) = nA + \sum_{i=1}^{n} \left(x_i^2 - A\cos(2\pi x_i) \right) \quad \text{with } A = 10$$

This is the reason why a search in the n-dimensional resulting landscape is a separable problem, meaning that consecutively solving the 1D global minimum search for one axis after the other while keeping the positions along the other coordinate axes constant (anywhere) will lead to successfully finding the overall global optimum. In the plotted 2D case this means twice in a row one has to find the right valley among seven available ones, whereas searching the whole 2D plane requires finding

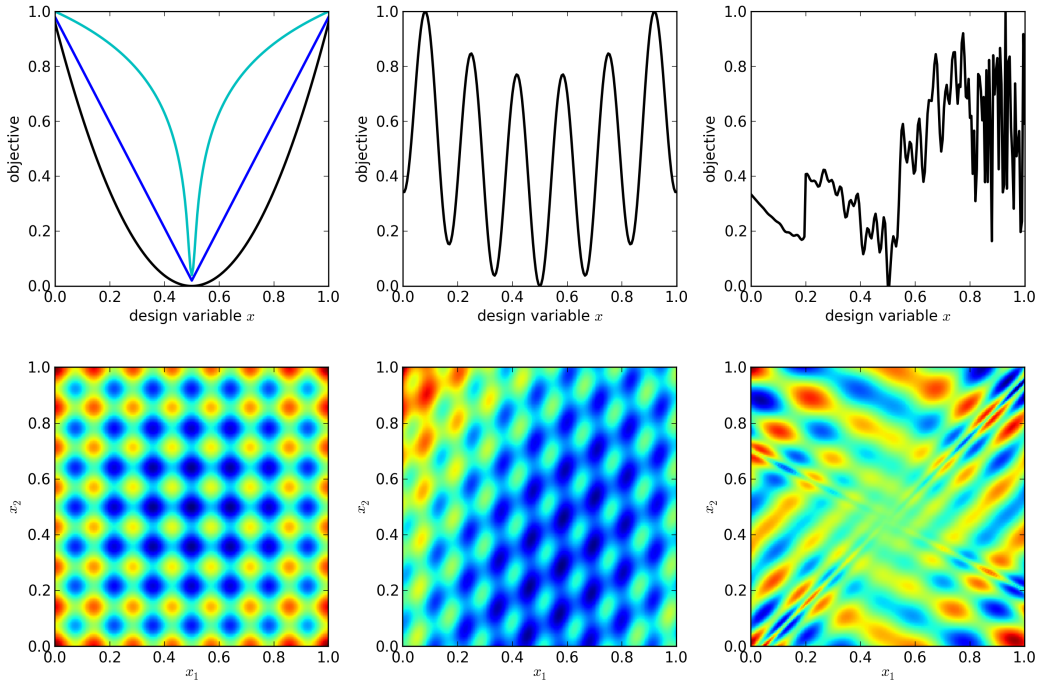


Figure T.4 Search spaces illustrating various difficulty levels.
The top row shows objective functions of one-dimensional optimisation problems, the lower row two-dimensional objective function landscapes. In both cases the search difficulty level increases from left to right. The common "jet" colour scale is used where blue stands for low and red for high numbers.

the deepest valley among 49 options. In order to illustrate how the problem size grows with dimensionality, here are the numbers for the 10D search: if the problem is separable then ten times isolating the right valley among seven will do, but if the problem is nonseparable, then there are 7^{10} valleys in the landscape that may have to be checked during a search, that's more than 282 millions.

The plot in the centre illustrates that a simple rotation turns a separable problem into a nonseparable one. On the basis of this added difficulty the rotated Rastrigin function is often used for benchmarking of minimisation algorithms. The transformation matrix applied here contains also a stretching component accompanying the pure rotation which is a common practice in test problem design [195, 265, 363]. However, it is clear that the other advantageous features turning the Rastrigin function into an easy problem are still there. The rotation does not destroy the regular pattern of the valleys. The regularity of the ripple pattern is a property of the landscape that can be exploited to efficiently search the global minimum even in the nonseparable case. The landscape to the right, by contrast, has a less regular valley pattern. It was invented and named function f_{101} by D. Whitley et al. [508, 511] who designed it to pose a more challenging search problem by avoiding any average gradient information. Think again of 282 million valleys. If they form a regular pattern, then they do not pose a big problem, but if they are so unregular[7]

[7]There is one thought giving hope that the degree of unregularliness of millions of valleys is

that many of them need to be properly characterised in order to come to a reliable conclusion about the location of the global minimum, then it looks quite different. Furthermore, it is also easy to think of deceptive landscapes like millions of valleys with a regular structure and one oddly narrow and deep one forming the global minimum which is hidden in a place far away from where the regular structure leads to.

Thus, the difficulty of a search problem in $\mathbb{R}^n$ hinges on:

- the dimensionality n,
- the presence of local minima,
- whether the problem is separable or nonseparable,
- the scale of structure (relative to the domain width),
- the type of structure, i. e. wavy, sharp edges, or cliffs,
- the degree of valley separation, i. e. whether smaller valleys open up into bigger ones or whether they are separated by barriers,
- the level of irregularity of the structure, and
- the level of noise.

If several of the disadvantageous properties come together, a search problem can get so difficult that it becomes impossible to devise an algorithm with which the global minimum can be found with a decent probability during a human lifetime, even with the best computer equipment. A mathematically more useful definition is that of a problem being *NP-hard*. This term from complexity theory defines the class of those problems that are *computationally intractable in polynomial time* [156], i. e. the computation cost grows with problem size n not in a polynomial manner (e. g. n^2 or n^5), but at exponential rates (e. g. 2^n or 5^n). Many real-world optimisation problems are known to be NP-hard; thinking of the exponential growth of the number of valleys in the Rastrigin function or f_{101} can make that plausible.

The curse of dimensionality

The discussion of the Rastrigin function above was used to exemplify a problem where the problem size grows exponentially with the problem dimension: in 1D, 2D, and 10D the landscape contains 7, 49, and 282×10^6 local minima, respectively. If the valleys show a simple regular depth distribution, then this is in principle no big deal, but in a less regular landscape where the global minimum can only be reliably found after having characterised a substantial part or all of the valleys, the huge computation cost makes anything than the few smallest problem instances impossible to solve under the blind search paradigm. This is the *curse of dimensionality*. The

bounded: it is the finite information content of real-world problem descriptions. If the problem can be described in a few kilobytes of source code, then there is no way to hide the information for individual depths and shapes of 10^8 valleys within. This is described by the *compressibility* of a function, see [405, 441] and section T.1.4.

problem becomes even more apparent after realising that characterising a valley means a lot more than finding it. One added sample can be enough for revealing the existence of a valley. But in order to say that the examined valley does not hide the global minimum, one has to scatter probing points over it along each dimension so that they lie densely enough in comparison to the underlying ruggedness to allow that claim.

The images in figures T.5 and T.6 are intended to give a feeling about the relations between problem dimension, structure scale, probing density, and probing cost. The top row of figure T.5 shows how both, a systematic (a) and random (b) search strategy in a 2D search space can reveal a coherent structure to the interpreting human. In both cases there are 144 tested points shown symbolically as either blue or red dots. Our ability of interpreting the landscape depends on the density of tested points, i. e. all over the field the *closeness* of test points has to be smaller than the length scale of the structure to be revealed. One could say that only the local existence of subsets of closely neighbouring points allows any interpretation. The second row with the finer structure illuminates this point because the sparse scans cannot reveal the structure any more. Unfortunately, for search spaces of increasing dimensionality, the conflict between the two goals of keeping the search global and at the same time establishing such local subsets of points becomes more and more severe.

The diagrams in figure T.6 are an attempt to further illustrate that point. Plot (a) shows the probability distribution of the distance between two randomly chosen points inside an n-dimensional unit cube. One can see that with increasing dimension the bulk of distances found shifts towards the length of the cube's longest diagonal. More importantly, given any two randomly chosen points, the probability of them being close neighbours reduces drastically as the dimensionality of their environment grows, implying that for higher-dimensional optimisation problems, when creating the test points randomly and uniformly over all the search volume, the establishment of locally close subsets becomes ever more costly. Based on the same statistics, plot (b) shows the percentage of cases where the distance between the two points is larger than 1, the side length of the cube, which illustrates as well the rate at which close point pairs become rarer with increasing dimensionality.

Assuming that by "challenging parameter search problem" we mean that there are more than, say, 4 dimensions, more than 3 or 4 valleys in each direction, where the valleys are not known to have a regular kind of spatial and depth distributions, and that one call to the objective function $f_{\mathrm{obj}}(\vec{x})$ costs more than just a few processor clock beats, the above thoughts should make it clear that in this case it is by principle impossible to examine the search space *closely* everywhere, and that algorithms will have to be used for the search which reject large parts of the search space after *merely rough* examination and save part of the limited computation budget for the few regions identified by the algorithm as being interesting. The realisation that a problem is NP-hard, i. e. computationally intractable, means that the hope of ever finding the global minimum and being sure about it has to be given up. This doesn't imply the global minimum will never be found. If the landscape contains some guiding features and the search algorithm is susceptible to them, then it may well be found more or less often. But if there is no way of probing the whole search

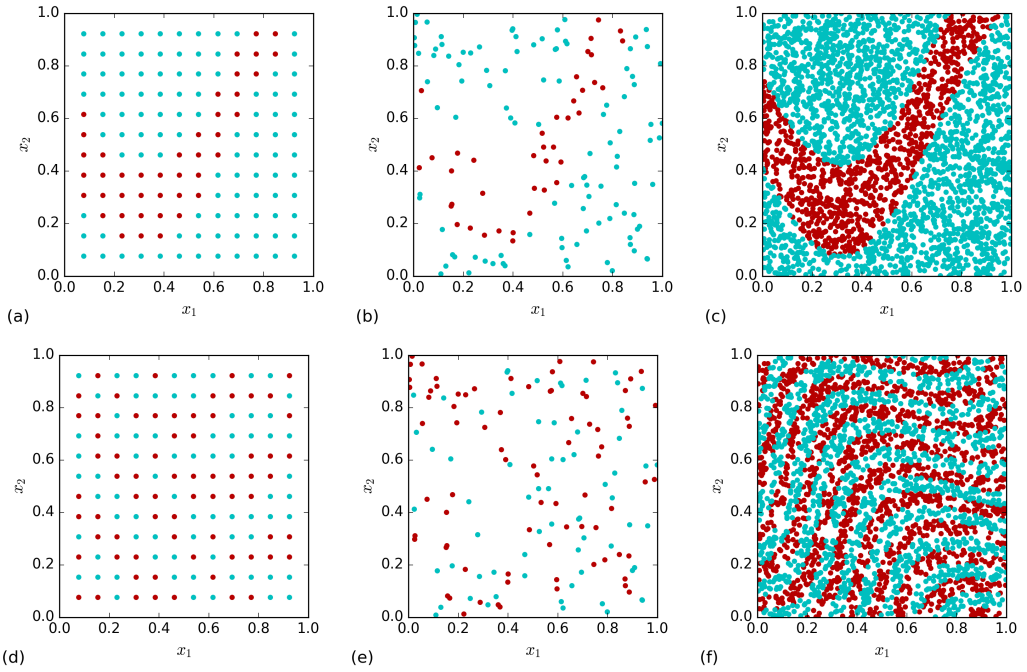


Figure T.5 Probing strategies: grid versus random
Points in the 2D search space $[0, 1] \times [0, 1]$ are being probed and yield either a value below threshold (indicated by blue colour) or a value above (red). Two different objective functions are probed in the top and bottom row. The plots show the results of two different probing strategies, a systematic grid (left) and random locations (centre and right). In each plot the number of tested points is 144, except in the plots on the right where it is 3000. In the top row $f_{\text{obj}}(\vec{x})$ reveals a structure with a characteristic length scale larger than the testing grid on the left. In the bottom row the length scale is comparable to the systematic grid. As a result, the true structure is only revealed in the densely probed case (f). Due to the comparable scales of structure and probing grid in case (d), aliasing leads to fields seeming mainly blue and others seeming mainly red, but the true average values within the apparent fields are all the same. The comparison between (d) and (e) shows the interesting case where a "lazy" random probing strategy helps to generate a data set less prone to misinterpretation than the data generated from a well-organised systematic exploration campaign.

The potential of pattern misinterpretation is not the biggest problem of regular probing grids. When examining the response of a system under variation of several input variables, the greatest disadvantage of a so-called *full factorial* design of experiments (DOE) is the eventual collapse along the dimensions. Imagine scanning a response f in a three-dimensional space $f(\vec{x}) : \mathbb{R}^3 \to \mathbb{R}$ employing a full factorial $3 \times 3 \times 3$ DOE. If one of the three input variables x_i turns out to be of minor influence, then there are triplets of experiments with very similar results because they differ only in the setting of the parameter x_i which has a negligible impact. The 3D probing grid of 27 points collapses along one axis into a 2D grid of 9 information-bearing data points. When an explorative scan, i. e. a sensitivity analysis, is conducted at the outset of addressing an optimisation task, then the information about which parameters have a heavy and which ones a minor impact is sought as one of the answers, this knowledge is assumed to be normally not available upfront. The a posteriori dimension collapse is the main disadvantage of a regular axis-aligned probing grid. In a full factorial DOE with three points along each axis, if 1, 2, 3, or 4 parameters turn out to have little impact, then 66, 89, 96 and 98.8 % of probing points are wasted. This is why randomised approaches like Monte Carlo samplings or Latin hypercube schemes are often highly advantageous.

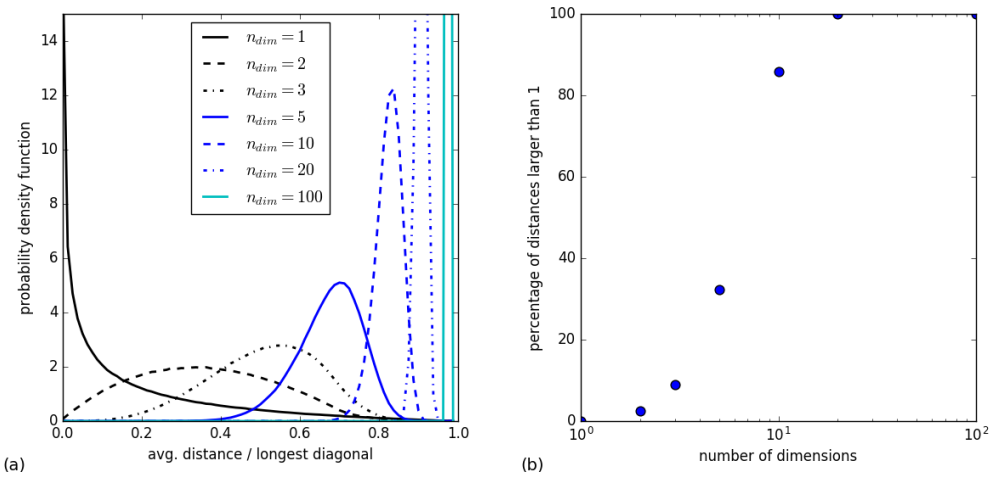


Figure T.6 The curse of dimensionality
These two plots illustrate the distribution of distances between a pair of randomly chosen points inside an n-dimensional unit cube (simulation of 4×10^5 pairs). Diagram (a) shows histograms approximating the probability density functions. Diagram (b) shows the probability that the two random points are farther apart than the cube's edge length.

space to come to a reliable mapping, then there will be no way of excluding the existence of oddly (narrow and) deep valleys that may get missed during some of the searches. Anyway, if they are missed in "blind" stochastic black-box search on a real-world problem in a case where there are no analytical means available offering a less blind search, then they will simply stay undiscovered and for all practical purposes irrelevant until tracked down by a better or more lucky search.

No free lunch theorems for EA

EAs are often described as a good first choice for tackling challenging real-world optimisation problems. But a certain sort of exaggeration of this sentence, has led to some rigorous theoretical work, resulting in a couple of so-called *no free lunch theorems* (NFL or NFLT). These theorems point out the danger that the notion of the *efficiency of EAs as general-purpose black-box optimisers* should not be misunderstood as a mathematically strict statement in the form of a claim that any particular EA might be *be significantly better than random search on no matter what type of optimisation problem*. We will see that in the case of a restriction to engineering-type optimisation problems under certain circumstances the notion is perhaps not that wrong. But the generalisation without disclaimer message certainly is. An early treatment of the question dedicated to combinatorial (discrete) optimisation with a GA was presented in 1991 by Hart & Belew [198], and Whitley [512] points to an even earlier NFL observation by Rawlins [377], but the probably most influential paper came out in 1997 and was written by Wolpert & Macready [519].

The arguments of Wolpert & Macready build upon the countability and finite number of search space vectors $\vec{x}$ in $\mathcal{X}$ and their objective function values $f(\vec{x})$ in $\mathcal{Y}$ in combinatorial optimisation problems. Let us pick the TSP as an example, where the task is to find the shortest route connecting n cities. There is a fixed

number ν of possible tours, represented by a list $\vec{x}$, and each tour has a total length $f(\vec{x})$. The question is in what order an algorithm goes through all these options and how many trials it takes until the global minimum is hit (or a quality threshold is undercut). The only relevant difference between two competing algorithms is thus how early it happens in each one. Not counting revisits of identical points, and assuming each algorithm is set up in a way so that it will eventually have checked all elements of $\mathcal{X}$, these are the two assumptions necessary to proceed in that argumentation. The NFL theorem can be proven by taking random search as one of the two competing algorithms and asking whether any algorithm can be more efficient than it on average over all thinkable objective functions f projecting from $\mathcal{X}$ to $\mathcal{Y}$. The result is that all algorithms – including random search – turn out to yield the same average performance. That result can be made plausible by scrutinising what can be hidden behind the term *all thinkable objective functions*.

Culberson's *two-men-on-a-park-bench* view of blind search is very suitable in this respect because it allows him to make the point that in blind search the person giving the answers might be an adversary making up the objective function on the go. He might decide to save up the best element of $\mathcal{Y}$ for the end no matter how much effort the other person puts into the sophistication of the search strategy. *All thinkable objective functions* gives the adversary absolute freedom. He might give just random values or he might lure the blind searcher into the impression that the algorithm worked out during the first half of the time only to give worse and worse data from then on.

This illustration can help understand the underlying thought of the article by Wolpert & Macready [519]: all the features that can help guide an algorithm into the right direction in one landscape will have a negative effect in one or more other thinkable landscapes. In the case of discrete optimisation the sets $\mathcal{X}$ and $\mathcal{Y}$ are countably finite, and the number of possible mappings f is therefore also countable and finite. The authors show that going through the whole set of possible functions f with a given algorithm can be translated into going through all possible permutations of the sequence of checking the elements of $\mathcal{X}$. The distribution of the moments of discovery of the global optimum $\vec{x}_{\text{opt}}$ is shown to be uniform and the averaged moment of discovery is the average position of $\vec{x}_{\text{opt}}$ in these permutations, which is of course the centre position. The clou is: applying random search to the whole set of mappings f will lead to the same uniform distribution of moments of discovery of the global optimum, and systematic enumeration will as well.

Thus, the NFL theorem can be stated as follows: *strictly considering all possible mappings f for given sets $\mathcal{X}$ and $\mathcal{Y}$ as target to a given search algorithm, there will be many versions of f leading to $\vec{x}_{\text{opt}}$ being checked out sooner or later by the algorithm, and in the end the average discovery time will be the same as the one seen by random search. Similarly, if instead of the discovery of the global minimum the hitting of a threshold value of $f(\vec{x})$ is sought, all discovery time distributions will be shifted but still equal between all algorithms.* This is a quite pessimistic result because it tells us that if it can be argued that a given optimisation problem corresponds to blind search, that there is absolutely no reason to expect any specific search algorithm to be better than random search. This seems to render all talk about *generic black-box optimisers* pointless.

Engineering problems and the NFL theorem

If we narrow the definition of engineering optimisation problems to continuous-domain parameter-tuning problems, where the parameters, in the spirit of the examples mentioned above, may stand for structure dimensions, angles, weights, material parameters, capacities, resistivities, inductances, parameters of transfer functions of system components etc., then we expect that small variations of $\vec{x}$ lead in most cases to changes in the objective function $f(\vec{x})$ which are also small. Here, "in most cases" means that we will not be surprised if some systems undergo phase or state changes causing cliffs in $f(\vec{x})$ but the function should nevertheless be piecewise smooth, or if there is noise, we assume there is a significant and piecewise smooth signal below the noise, otherwise we would have engineered a different $f(\vec{x})$. This means that the hypothetical adversary in Culberson's *two-men-on-a-park-bench* model loses a drastic part of his freedom. The loss of the adversary's power connected to the justified assumption of piecewise smoothness translates for us at least into the ability to conduct useful local searches, i. e. following the local gradient by conducting a descent in the steepest direction, if necessary with a noise filter. But how about global search? For global search strategies to be useful there must be overarching structures spanning large areas of the search space. If such structures are not present, if there are no long-distance correlations, if size, shape, and depth of all the valleys and size, shape, and height of all the mountains are all individual and there are no overarching properties, then this means nothing else than an NFL on the level of valleys, and all valleys have to be characterised individually, because from collected data on a subset of valleys there is no information on any of the remaining valleys. In this case there is no better algorithm than local search with random restarts. As a thought experiment, the local search could be taken away from the search algorithm and be made part of the objective function. Then the search landscape turns into horizontal plateaus at different heights and with varying sizes and shapes. One ends up with a coarsened random landscape, and on this coarsened level there is no free lunch again. Since the plateau heights are distributed randomly, there can be no better algorithm than random search or systematic scanning by enumeration.

Luckily, two other facts can be assumed to further diminish the freedom of the adversary: the underlying physics of the system and the compressibility of the function. Regarding the first point it can be said that the more complex the analysed system is, the more complex structures may show up in the objective function. In different areas of the search space the system may be in different working regimes. But it still is the same system, and throughout the subspace covered by one working regime there should be one, more or less easy to interpret, pattern. The second point deals with the compressibility of a function. Imagine again the adversary making up randomly distributed objective function values on the fly to fake an incredibly difficult search space. At least he would have to note down his past output in order to be able to repeat the values consistently in case the searcher revisits a point or its close vicinity. If the adversary forgot the values, then he would in fact produce random numbers instead of faking a rugged landscape of fine resolution. The finer the structure and the higher the search space dimension, the larger will the amount

of data become which is necessary to describe the highly complex function.[8] The adversary would have to keep track of all that data. A finely resolved but highly regular landscape, by contrast, needs just a small description, like the formulae of the Rastrigin function or f_{101}. A function is *compressible* if less data is required to describe it than there are elements in $\mathcal{Y}$. If physical systems like circuit boards or steamships are to be optimised, then an ensemble of physical laws and a set of system component specifications will carry all the characteristic information of the system. The objective functions of real-world engineering problems are highly compressible. Even if large simulations of the fluid dynamics of a steamship have to be carried out over hours on many cores of a supercomputer to evaluate one design, the ship design description and the equations modelling the fluid physics will fit on one old-school magnetic disc.

Therefore, the system's physics and the compressibility [185, 219, 405, 441] (and references in [185]) of the problem description which can be assumed to lead to some extent of regular pattern, together with the basic mathematical measures of well-behavedness [159][9] of $f(\vec{x})$ (like e. g. piecewise smoothness), these are the facts upon which the escape [19] from the no free lunch theorems hinges. This is what allows leverage for optimisation algorithms designed to be general-purpose blind search machines for parameter-tuning in engineering problems.

T.1.5 Summary

The examples of rope and boat optimisation were to show that the objective function itself is a decisive ingredient in the numerical optimisation process. The objective function can be modified superficially by the introduction of penalties, and it can be fine-tuned in depth, as to change the amount, relative depth, and width of valleys or the height of separating barriers in order to ever more reliably and efficiently guiding the applied search algorithms to desirable solutions. With the curse of dimensionality and the no free lunch theorem in mind, the importance of not neglecting available ways of tuning the search-topology should become apparent.

Objective functions are not chosen, they have to be engineered. They comprise a considerable part of the problem solving process. They can be an important determinant of the cost or efficiency of a numerical optimisation. Numerical optimisation means shifting the low-level part of the work of solving an engineering problem to computers. The automation enables a large increase in quantity of the low-level work. As a consequence, high-level work arises on the conceptual level, the vast research literature on EAs and other optimisers bears witness of it. But the person intending to apply the fruits of that conceptual work should keep in mind that there is still another part of work remaining, which has to be redone with each new application case of different nature, the work of thinking about the implementation modalities of the objective function.

[8] Another illustrative thought experiment highlighting this point can be found in [219] on pages 53 & 54.

[9] Gaviano & Lera write: "[...] we may know the Lipschitz constant, a bound on the second derivatives, the number of local minima, a bound on the size of the 'region of attraction' of each local minima. Clearly, in such a case, we expect that algorithms that exploit these additional properties, have better convergence properties."

Evolutionary algorithms are not needed in really simple search landscapes. When there are just a few valleys, in those cases gradient search with a couple of random starting locations will do. And when properties are known making the search space very regular, like e. g. the rotated Rastrigin function, or giving it another similarly exploitable peculiarity, then a specialised search strategy can easily be devised. Evolutionary algorithms are interesting for the "seriously difficult searches" remaining. The ingredients making a landscape difficult are local minima, the curse of dimensionality, irregularity of structure, and small scale of structure. Any optimisation problem ending up in the set of "seriously difficult searches" is there either because f_{obj} is known a priori to contain a subset of the notorious features or because other optimising strategies fail (e. g. when they repeatedly stall in different local minima, this can reveal a posteriori that the unknown search landscape must have such features). In many simulation-based engineering problems it is the default case not to know very much beforehand about the topology of f_{obj}.

EAs are heavily based on the random number generator. Basing a search engine on random distributions is a safety and robustness measure. The advantage can be expressed in two ways: probing patterns in the form of random distributions are not prone to misinterpreting regular structures, and the efficiency of algorithms employing them can be assumed to show a lowered susceptibility to changed problem topologies. This translates into an increased reliability when going from training problems to the unknown objective function landscape of the real problem of interest.

EAs are worth a shot when it can be deemed that the structure of f_{obj} contains a difficult mixture of exploitable and blurring or deceiving features. EAs try to filter out the useful information content and ignore the barriers or rugged areas in the landscape that block the progress of many deterministic search algorithms. They do that based on different combinations of search heuristics. Hence, some EAs are more suitable than others for certain search landscapes. Thoughts on no free lunch theorems imply that in the case of continuous-domain parameter-tuning problems with real-world engineering background where a certain level of well-behavedness can be expected from f_{obj} most of the time (in contrast to discrete combinatorial problems) there is a fair chance that there are some useful features in the search landscape, and EAs are designed as conceptually simple but robust methods to capture some of it. It could be said that EAs are either a tool of the desperate or the lucky one, either it is the last thing you try before random search, or it can be the joker helping to avoid the work of thinking about and modelling a problem.

T.2 EA terminology

In the field of EC the languages of mathematics and computer science are enriched by a lot of terms from biology which offer convenient shorthands for many useful entities and operations. It starts with the solution candidate $\vec{x} \in \mathbb{R}^n$, the subject of optimisation through parameter tuning, which can also be called chromosome, genome, or individual. The three terms don't overlap in biology, but they do here depending on what particular abstractions have been employed to distil the EA. The components of $\vec{x}$, the parameters x_i, are the genes. The objective function $f_{\text{obj}}(\vec{x}) : \mathcal{X} \to \mathcal{Y}$ can also be called fitness function or simply fitness. Greater fitness

is always better, so in a minimisation problem $f_{\text{obj}}(\vec{x})$ and the fitness are not quite the same. Often the tunable parameters are bounded, i. e. $x_i \in [a_i, b_i]$. EAs act on populations of solution candidates $\mathcal{P} = \{\vec{x}_j\} = (\vec{x}_1, \vec{x}_2, ..., \vec{x}_N)$. Let $\mathcal{P}_g$ be the status of population $\mathcal{P}$ in generation g where $g \in [0, 1, ..., G]$. The overall scheme of a generic EA is:

Algorithm 3: generic generation loop of an EA

```
P ⟵ GenerateRandomPopulation(N)
f_obj ⟵ Zeros(N)
while g <= G and not StopCriterion() do
    for j ⟵ 1 to N do
        f_obj[j] ⟵ Evaluate(P[j])
    P ⟵ SelectionMutationRecombination(P, f_obj)
```

An EA is mostly determined by what happens in the last line of the algorithm, where the population is being replaced by a new set of chromosomes which are generated based on the information from the evaluated old set. The construction of new chromosomes can be based on all kinds of mathematical recipes, but with the inspiration from biology, *selection*, *mutation*, and *recombination* routines have often been invented to act on the set of chromosomes $\{\vec{x}_j\}$ and create a new set $\{\vec{x}'_j\}$. The goal is that from generation to generation more and more beneficial parameter combinations are discovered, become fine-tuned, and accumulate in the gene pool. This is sketched in figure T.7. For this to happen, the algorithm must find a good balance between *exploration* and *exploitation* (or *diversification* and *intensification*), this means the limited budget of calls to the objective function has to be allocated efficiently serving the two opposing purposes of probing yet unexplored regions of the search space and searching more densely in the most promising subspace covered by the best members of the population. Often, mutation can be seen as serving the purpose of diversifying the gene pool whereas increased selection pressure and recombination are the means of intensification. Seen from a more abstract level, mutation and recombination are just operators creating new chromosomes from old ones, either as a unary operator using just one chromosome, or as binary, ternary, or higher-order operators using two, three, or more of them.

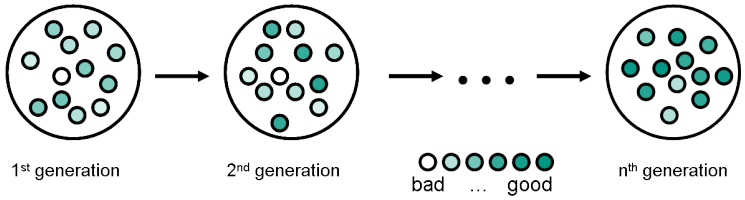


Figure T.7 General generational scheme of an EA.
A population of candidate solutions is being optimised in an iterative cycle of generations. Each dot symbolises a candidate solution and its fitness is shown by colour. An EA is defined by how exactly selection pressure and the operators modifying parent chromosomes are implemented. By principle, everything is allowed. The fitness evaluation procedure is treated as a black-box by the algorithm. The only data the EA works with are the chromosomes $\vec{x}_j$ and scores $f_{\text{obj}}(\vec{x}_j)$ of already evaluated individuals.

Selection pressure is the implementation of different probabilities of individuals to get (all or part of) their chromosome reproduced in the next generation. This can be done either based directly on the fitness values or on the ranking within the population. The chromosomes can have individual reproduction probabilities between 0 and 1 or they can be divided into groups with equal probabilities among group members, where the simplest case is two groups with reproduction probabilities 1 and 0.

The case where only a subset of size $\mu < N$ from the whole population of size $\lambda = N$ "survives" and can reproduce is commonly described as a (μ, λ)-EA. The useful μ-λ-notation became widely adopted after having originated from the literature on evolution strategies (ES). By convention, μ is the number of parents, λ is the number of offspring, and γ can be used as the number of generations. Instead of the comma which stands for a cutoff in the parent population, the symbol "+" can be used to describe a union of the two chromosome pools which implies a survival cutoff in the resulting set. Thus, a $(\mu + \lambda)^\gamma$-EA means λ offspring chromosomes are created from μ parents, and from the union of the two sets across the two generations, the μ best are selected to become the new parents. This goes on for γ generations. The notation can also be generalised to describe simple patterns of separating and merging populations [380].

Since an EA acts only on the representation $\vec{x}$ of a solution to a specific problem, and often there are various ways of representing the search space, it makes sense to borrow two more words from biology: *genotype* and *phenotype*. If the parameter list $\vec{x}$ is $(8, 2, 13)$, then it can be represented as $(8.0, 2.0, 13.0)$, $(1000, 0010, 1101)$, (100000101101), or, transforming $[0, 20]^3$ into $[-1, 1]^3$ it could be $(-0.2, -0.8, 0.3)$. These are all different genotypes for one phenotype. The various types of genotype representations offer different ways of implementing the chromosome-modifying operators. Mutation can mean either adding a number like 0.1 to one x_i or flipping one bit in a binary representation. The EA acts only on genotypes, whereas the experimenting human is interested only in the phenotype, i. e. the "worldly" appearance of the represented solution after decoding the chromosome (e. g. the particular instance of a parametrised computer program or construction plan). Maybe one is intending to tune two masses measured in kilograms and a rope length in metre, and finds out that different EAs with different genotype representations show different efficiencies at finding a good solution. Judging a final solution is often only possible after transforming the list of numbers back into a simulation of the problem and looking at a specific visualisation.

T.2.1 EA $\subset$ metaheuristics

Metaheuristics (MH) should be understood as heuristics of heuristics, where the noun *heuristic* means a building block of algorithms which can be a subroutine or a minor algorithm of its own, but sometimes single operations can be meant. In this view, the coinage of the term *metaheuristic* by Fred Glover [169] and his inclusion of canonical genetic algorithms into the definition can be easily understood.[10] The

[10]Whereas *heuristic* generally means *ars inveniendi*, its meaning in the research literature on algorithms for discrete combinatorial optimisation problems must have narrowed down to *a routine*

meaning of MH can be analysed very clearly with a simple example by going back to Glover's original reason for including the simplest GAs. Take a $(1, 2)$-EA where offspring is only generated by taking the single parent chromosome and mutating it. In a $\mathbb{R}^n$ representation, mutation means adding a number, and in a binary representation, mutation is flipping a bit. In that case the mutation operator is representation-dependent, but the EA is not. The two versions of the mutation operator are called two low-level heuristics. The problem-independent higher-level part of the EA is a metaheuristic.

More definitions and literature can be found in [54, 75, 117, 162, 280, 283, 338, 340, 452, 493, 514, 534]. In the above example the higher-level metaheuristic is not very complex and not at all adaptive. This almost violates more recent definitions[11] like the one by Caserta & Voß [75] who stress the distinction between a *guiding process* and the *application process* or Osman & Laporte [338] who go even further:

> "A metaheuristic is formally defined as an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting the search space, learning strategies are used to structure information in order to find efficiently near-optimal solutions."

This means the guiding process must be responsive to the currently experienced search landscape. It is a much stronger definition where the above simple EA example or canonical GAs are definitely excluded. Due to the stress on the notion of responsiveness, this definition will in many cases overlap with the definition of memetic algorithms[12] (MA), in particular, an EA where the responsive flexibility is related to the tuning of a local search subroutine will equally fall under both categories. There is the difference that in MHs the adaptation feedback cycle is supposed to rest in the higher-level part of the algorithm whereas in MAs it is thought to be a feature emerging from additional functions being properties of what is an individual to the higher-level EA. MH and MA can be seen as two different conceptual thinking modes that can (but need not) lead to algorithms indistinguishable on the implementation level.

Going with the broad initial definition by Glover [169] which is also reflected in an encyclopaedic definition by Dorigo et al. [117], metaheuristics is the broader term. In that view the nature-inspired inventors of historic EAs ended up with special instances of MHs. Given the fact that up until the 1990s the schools of GA, ES, etc. existed rather separately from each other and from other streams like tabu search or simulated annealing, one can say that the coinage of the broader terms EA and MH happened late enough but then helped to clarify the conceptual similarities and differences and open up new connections.

---

consisting of several basic mathematical operations whose invention and usefulness are not based on stringent mathematical analysis and proof, but rather on intuition, common sense, and empirical results. This can be seen from Fred Glover bringing it to the point very succinctly in [171]: "Algorithms are conceived in analytic purity in the high citadels of academic research, heuristics are midwifed by expediency in the dark corners of the practitioner's lair."

[11] Zäpfel et al. compare a few definitions in [534] on page 72.

[12] see section T.4.9, p. 555

T.2.2 Elements of the modern theory of evolution

> "Darwinian evolution is no more than the inevitable consequence of competing information-reproducing systems operating within a finite arena in a positively entropic universe."
> (W. Atmar [15])

The theory of evolution was developed at a time when old knowledge of breeding dogs and horses was put in connection with new discoveries like skeletons of prehistoric animals and systematic similarities in many animals' skeletons. That was long before the discovery of atoms and molecules, and the microscopic layouts of viruses and bacteria. The discoveries of molecular biology and other modern branches of biology have ever since reinforced Charles Darwin's fundamental ideas on the origin of species. But of course, the theory became refined in a lot of aspects after incorporating the new facts. This history has led to naming today's more complex theory building the *modern synthetic*[13] *theory of evolution*. In the following, the key concepts will be recalled by dividing them up into the level of individuals and the level of populations. On the level of individuals they are:

- **gene code:** The sequence of nucleic aids (adenine, cytosine, guanine, thymine) composing a DNA molecule has a meaning, there is an alphabet. The single letters are formed by triplets of nucleic acids, there are triplets coding for starts and stops of genes, the other triplets code for amino acids. The DNA molecule exists as a double helix, it can be split into two single threads with little energy. Single threads can serve as copy masters for synthesising the complementary thread. RNA copies can be made of DNA sequences. *Ribosomes* act as sites where proteins are synthesised by forming chains of amino acids according to the triplet letters of a gene. Complex organisms have their construction plan spread across several DNA chains, called chromosomes. The meaningful segments of text are called genes and they are scattered along the chromosomes. Genes are meaningful units because each gene codes for one protein. Different versions of a gene occurring in the gene pool of a population are called alleles.

- **inheritance:** The DNA of an individual is basically a copy of the DNA of its ancestors, apart from small random modifications. A modified genome is passed to the offspring only when the modification occurs in the germ cells (*gametes*).

- **mutation:** Modifications of the genome are called mutations. There are point mutation (modification of single nucleic acids), deletion, insertion, and mutation types on the scale of chromosomes. Insertion mutations create redundant pieces of genetic information. One of the copies can be used for tinkering around, and the corresponding protein may stay more or less useless for many

[13] The word "synthesis" (from Greek σύνθεσις) is generally used to describe a merging of components leading to a new quality. Here it stands for the new quality, which the theory of evolution gained after knowledge of the microscopic world (existence of chromosomes) and later from molecular biology (deciphering of the genetic code; discovery of how DNA is translated into proteins by the help of mRNA, tRNA, and ribosomes; relating genes to phenotypes along research on the fruit fly drosophila melanogaster) has been incorporated.

generations. Such useless but also harmless genes stay present in small fractions of the gene pool and are free to mutate further until the emergence of a very different new functionality may gain weight in improving individual fitness.

- **gene recombination during reproduction:** Bacteria reproduce simply by cell division. Every bacterium has just one parent and is an identical copy of it apart from possible small mutations. But all organisms of higher complexity reproduce by combining genetic material from two parents. What is the advantage of this? Imagine two mutations which increase fitness only in conjunction and which are neutral alone. Without incorporating DNA from a second parent, in order to combine the two mutations in one line of descendants they both have to actually happen in series along that chain. This corresponds to the multiplication of two tiny probabilities. A delay of many generations may exist between the two mutations of interest and a lot of other mutations can happen before. Because of the game of probabilities, the two mutations might show up in separate lines of descendants, but only with greatest luck in one single line. By introducing the concept of forming offspring by the combination of two parents' genomes (sexual reproduction) the probability game changes drastically. In such a regime the probability for two neutral mutations occurring at low frequenceis in a population's gene pool of getting combined in one individual is much higher. Once the conjunction is present in one line of descendants, it can play out its fitness advantage, and this acts as a force increasing the frequency of the combination in the population's gene pool. Without recombination the graph of lines of descendants consists only of branches and looks like a tree. With recombination it looks like a net. Because of the game of probabilities the EA with recombination is much more efficient at improving the gene library of a species than the EA without recombination. One of the most primitive life forms with systematic recombination of complete genomes of two parents is green algae[14]. The invention of the recombination operator (and with it death[15]) greatly boosted the speed of Mother Nature's evolutionary algorithm. More complex life forms could only develop because of the thus accelerated EA.[16]

- **haploid and diploid cells:** More complex life forms are built with diploid body cells containing double sets of chromosomes. In the case of humans it is 23 pairs of homologous chromosomes. Germs cells (gametes) are produced by *meiosis*, the process of cell division whereby diploid cells are divided into two haploid ones by randomly dividing up the homologous pairs of chromosomes. New individuals are begotten by the fusion of two haploid gametes. The systematic random remixing of chromosomes in each generation is the

[14]Two haploid green algae cells can merge and form diploid zygotes, this is called conjugation [271]. (see also `http://en.wikipedia.org/wiki/Green_algae`)

[15]Why was the invention of death beneficial for Mother Nature's evolutionary algorithm? Appendix U.1 adds a few thoughts on this.

[16]"Of course" some form of reconnecting lineage branches has also emerged among bacteria, see appendix U.2.

top level recombination operator in nature's EA for complex life forms. The redundancy of the genetic code in diploid cells leads to a certain amount of freedom of experimentation with genetic code. Another consequence are the rules for the phenotypic expression of recessive, intermediate, and dominant alleles called *Mendelian inheritance*[17] discovered by Gregor Mendel in 1865 [296]. Exceptional uneven redistributions during meiosis are called genome mutations, they are the reason why the number of chromosome pairs varies between species. A doubled chromosome is another redundant copy of genetic code to experiment with.

- **crossing-over or crossover (CO):** CO is the next lower-level recombination operator of nature's EA for complex life forms. In the process of meiosis, before dividing up the diploid set of chromosomes into two haploid sets, pieces of chromosomes get systematically exchanged with a certain probability between the homologous pairs. The reason is the same as described above: without CO a combination of several beneficial mutations along one chromosome could only be found by consecutive mutations in one line of descendants, for which the probability is too low. The cut points along the chromosomes do not seem to be completely random-determined, there are hot spots [290]. Most of the time equivalent pieces are swapped. The rare events where this is not the case can serve as moments of enlarging or reducing the size of chromosomes. CO would pretty much always cut through and potentially destroy important genetic code if not large parts of "useless empty code" existed in between the expressed genes.[18] RNA splicing[19] is one more of nature's inventions, it enables recombining different functional parts of proteins by CO.

- **selection pressure:** This is meant by the popular term "survival of the fittest". It has to be kept in mind that "fit" does not always have to mean physically strong or fast or big, but rather fitting into the own environment so as to make efficient use of resources, balancing egoism in the fight for survival with altruism towards individuals with similar[20] genes or packs or herds with the ultimate goal[21] of giving numerous life to offspring in a way that the offspring will be able to do the same.

- **sexual selection:** Why are some non-dangerous and non-poisonous animals colourful and easy to spot? This does not help in evading predators. But it helps for mating purposes. This shows that the mechanics of evolution unfolds an own game theory. Not being seen is good with respect to predators, but not in terms of spreading genes. Therefore nature does not always seem to come up with the most energy-efficient engineering solution, as can be seen with

[17] `http://en.wikipedia.org/wiki/Mendelian_inheritance`

[18] Often findings of research in molecular biology explain older observations. Here is a good example of the old theory explaining a new finding, the discovery of "useless code".

[19] i. e. the scattering of more useless code sections within one single gene, see `http://en.wikipedia.org/wiki/RNA_splicing`

[20] normally only offspring, yet in the case of e. g. state insects also siblings

[21] It is understood that evolution is a random process without any goal (τέλος). All there is (or more precisely: all there can be explained by the scientific method) are emergent phenomena.

birds like peacocks or birds of paradise if regarded as flying machines. There are always multiple selection pressures at play, and the pressure to find and impress mating partners is just as decisive as the more obvious ones like the selection pressures created by predators or sparsity of food. The nightingale with unremarkable or even camouflage looks but very characteristic mating calls exemplifies an efficient way of responding to sexual selection pressure while not so much compromising on the safety from predators. By contrast, bird species where the mating competition is based on courtship rituals involving both ear-catching songs and eye-catching colours and behaviour prove that there can be contexts where sexual selection is a dominant factor over the long term.

But there is not only the level of individuals. The rules just described create a game theory of survival and reproduction. This leads to emergent phenomena on the level of populations which will be described below to help understand the basic mechanics of evolution.

- **gene pool:** This is the ensemble of genetic information held by the population. While one individual can have maximally two variants of the same gene in its genome, the gene pool of the population can hold many different versions. It can be seen as a storage facility for information. Genes with neutral or only slightly disadvantageous effect and recessive alleles can remain in the population in sparse distribution. When changes in the environment of the species or other mutations within the gene pool turn that variant suddenly into an advantage, it may begin to spread over most of the population within few generations. A diverse gene pool serves as a library, a safety net, storing information and enabling the population to respond more quickly to environmental changes.

- **genetic drift:** That's evolution's gradient search technique continuously looking for the local optimum. The peppered moth[22] is a famous example which has allowed to observe the phenomenon over the last 200 years. As tree barks in parts of England turned black with early industrialisation and cleared again later, the peppered moth could be observed to track the optimum with respect to camouflage. Selection pressure from birds picking those moths which are easier to spot drove the genetic drift. The frequency of the dark moth variant increased from 0.01 % to 98 % by 1895 and decreased again in modern times.

- **separation or isolation and speciation:** Genetic drift explains how species change their appearance over time, but it cannot explain how the number of species can increase (because recombination constantly homogenises a gene pool). In order for that to happen, the gene pool of one population has to be separated into two pools and the separation has to be kept up long enough until genetic drift has driven the two gene pools far enough apart so interbreeding becomes improbable and/or impossible when the populations are merged again (different ecological niche, look, mating call, infertility of mixed offspring (like

[22] `http://en.wikipedia.org/wiki/Peppered_moth`

mules) or other barriers to reproduction). It is believed that division onto the different islands of the Galápagos archipelago facilitated speciation for the Darwin finches. Also, researchers deem it probable that the development of fault grabens, mountain ranges, and lake systems in East Africa during the latest several million years and the regional climatic effects have played a role in the speciation of australopithecines because their fossils have only been found in East Africa [289].

- **population bottlenecks and founder effects:** A catastrophic sequence of events can diminish the number of living individuals of a species drastically. As one option, this may end up with the extinction of the species, but as an alternative option the population may grow again and recover. In the latter case the species goes through a so-called population bottleneck. Another occasion for population bottlenecks is the separation of a founder population, e. g. when a small group of birds is blown by a storm to a remote island, or when some insects or plant seeds are carried over a mountain range by a larger animal. The separation of founder populations can be assumed to play a major role in evolution and the creation of new species. During periods of reproducing in equilibrium in a stable ecological niche it can be assumed that selection pressure and gene drift often act as stabilising forces keeping the phenotype mostly constant (stasis) which implies little or selection-neutral changes in the gene pool. As a population bottleneck involves a different type of random selection as acting under stable conditions, and as the founder population may find itself in a different environment where it may be able to conquer a slightly different ecological niche, it may lead to a rapid growth of the new population if the new conditions are favourable. This is called a *founder flush*. With selection pressures acting differently in the new environment with other resources, enemies, and competitors, and with genetic drift leading the population towards a new equilibrium, the process of random-selected small founder populations going through founder flushes can be deemed to be a central driver of the evolution of species [177, 292, 293, 328, 366, 469, 470].[23]

- **ecological niche:** There are many different ecosystems in many different climate zones, but some structures, some techniques of gathering food, some symbiotic functions like pollenating flowers occur repeatedly and are called an ecological niche. For example, bumblebees and colibris both pollenate flowers and nourish on nectar. They share the same ecological niche and it drove the colibri to become (in size, wing motion, flight pattern, beak and tongue shape) very similar to the bumblebees.

- **higher-order phenomena: co-evolution, symbiosis, mimicry, etc.:** It can be assumed that many predator-prey pairs evolve in co-evolution following

[23]The view on evolution as a process controlled by the interplay of stasis in equilibria and innovation in connection with founder effects has led to the theory of *punctuated equilibria (PE)* [123, 177, 267] which may be seen as contrasting – or rather complementing [236] – older concepts of *phyletic gradualism*. PE can explain extended periods of phenotypic stasis and the "sudden" (in geological times) appearance of species in the fossil record.

the scheme that when rabbits become faster, only foxes that are able to follow the trend will survive. Co-evolution means that tightly interacting species determine each other's survival conditions in the ecological niche. Predator-prey pairs are examples of competitive co-evolution. Bees, bumblebees, moths, colibris etc. are in relations of cooperative co-evolution with flowering plants. The result of strong cooperative co-evolution with benefit to both sides can be symbiosis, e.g. lichens (consisting of fungi, responsible for protection and water supply and algae performing photosynthesis) or ants (protection) and plants lice (sugar). Mimicry is the phenomenon when non-dangerous plants or animals "try" to look like dangerous ones in order not to be eaten or attacked; it is therefore a tight co-evolution relationship.

Let us summarise some quintessences of evolution:

- **local search:** The interplay of mutation and selection works gradually. The probability of severe mutations to be beneficial is zero.[24] Only redundant gene sequences can mutate more freely. New information is generated gradually. Mutation and selection lead to genetic drift which is the local search tool of evolution. Many genes are steadily locally optimised by genetic drift in parallel. This is possible when they pose separable problems. (Consider these four separable problems of "online optimisation" of an eagle: the lens has to fit to the eye dimension, the wing size has to fit the body, the egg shell has to have the right thickness, the beak has to be in a useful shape.)

- **global search:** Redundant genes, speciation, founder effects, and long-term ecosystem variations are the important global search tools of evolution.

- **recombination:** Recombination leads to horizontal gene flow and acts as a cohesive force on the gene pool of a species, i. e. it 'keeps it together' (homogenisation as local search force). It greatly enhances the rate at which combinations of beneficial mutations are tried out in single individuals (explorative, i. e. global search force).

- **branching of code:** Evolution generates meaningful DNA literature. This shouldn't be seen as starting off with one long random DNA sequence and

[24] Mutations with severe effects on the system are meant here, independent of whether concerning the modification of just one nucleic acid or large genome parts. The reason why not to expect benefits from mutations with severe effects is the same as why one cannot expect from a bunch of bricks thrown up into the air to fall down and in place forming a bridge. A point mutation has a severe effect if a swapped amino acid changes the geometry and charge distribution in a functional part of an important protein drastically. That the changed geometry will efficiently fulfil without further tuning a different relevant enzymatic or structural task is as unlikely as randomly placed bricks forming a useful construction. (It can be reminded that proteins with central metabolistic functions (e. g. cytochromes) are like basic chemical gearwheels and appear in equal or very similar shape in all types of living beings, i. e. they are shared by archaebacteria and humans and almost every random amino acid change means a lethal mutation. By consequence, proteins have phylogenetic trees.) For mutations affecting large genome parts the argumentation should be even clearer. Either it is not severe (e. g. if it creates redundancy by copying or modifies text of which a backup exists) or it must by the laws of probability be disadvantageous, since constructive planning cannot be expected from a random process.

randomly changing letters until in blind search the code of a T-Rex is suddenly there. The microscopic details of cell biology reveal that it has more to do with the steady buildup of an ever growing and differentiating template-based code library by copying subroutines and incrementally modifying the redundant copies. Instead of one huge combinatorial problem there are many separable construction and fine-tuning problems.

- **branching of function:** Most pieces of active gene code in the library have a specific function at any point in time: on the way from dinosaurs to birds skin scales developing fringy structures and becoming more and more overlapping must have served the purpose of thermal insulation before they became useful as feathers for flying. The search is local, functionality is added incrementally.
- **The result:** Life is incredibly complex, many ecosystems arise, and innumerable species evolve in the interplay with each other and the planet.

The modern theory of evolution and the NFL theorem

In his article on the "futility of blind search" Culberson [98] asks whether the obvious accomplishments of Mother Nature's evolutionary algorithm (let's introduce the abbreviation MNEA) are at odds with the NFL theorem. He puts it in a nutshell by pointing at the combinatorial aspect of synthesising proteins which are just chains of proteins from the 21 types of amino acids common in eukaryotes. He cites a small calculation by Kauffman [225] noting that for a modest-sized protein of 100 amino acids there existed about[25] 10^{120} different amino acid combinations. This number is huge even when compared to the estimated number of atoms in the universe which is 10^{80}. Kauffman estimated that throughout the history of planet Earth, even under unrealistically advantageous assumptions, no more than $1/10^{60}$ of them would have been considered. But what we see in living cells is the interplay of thousands of proteins, each one highly optimised for its structural or enzymatic task. The impossibility of composing the genetic code for such a complex system by conducting blind search in the *two-men-on-a-park-bench* manner[26] led Culberson to look for ways to escape the implications of the NFL theorem and express these thoughts in a highly speculative last part of his paper. He proposes a way out of the dilemma putting forward three main arguments: (a) questioning whether genetic evolution really is efficient at solving complicated combinatorial optimisation problems, (b) delegating the invention of many chemical building blocks of life to the pre-life phase of evolution, and (c) questioning the blindness dogma at the level of chemical evolution[27], where the computation mechanism is not independent of the search environment. To support this line of arguments he refers to research by Kauffman [225, 226] on the tendencies of self-organisation of boolean networks "on

[25]The number of possible combinations to form a chain of 100 amino acids from 21 different types is $21^{120} \approx 1.67 \times 10^{132}$.

[26]see section T.1.4

[27]The term "chemical evolution" describes the accumulation of ever more complex organic molecules and cellular structures that is assumed to have happened during the evolution from pre-life to life (also: abiogenesis or biopoiesis). See `http://en.wikipedia.org/wiki/Origin_of_life` or the references, in particular 11-15, in the introductory part of [365].

the edge of chaos" and follows the suggestion that analogous tendencies of networks of chemical reaction pathways in the primordial ocean might have filled the gap left by an NFL-constrained EA to explain the high degree of orderly complexity of life as we see it. Furthermore, adding a reference to a paper containing debatable conclusions by DeJong [105] with the title "Genetic Algorithms Are NOT Function Optimizers", Culberson rounds up a view in which most of molecular engineering having made life possible must have happened before life and evolution as we know it had kicked off, and in which genetic evolution since then has been fulfilling functions of equilibrium maintenance and production of "only very localized advantage" in a "vast richness of opportunity in this highly structured universe", a view[28] in which genetic evolution had nothing much to do with what is conventionally seen as "optimisation".

This seems at least counter-intuitive. How can the shape of a spider web and the molecules used for its fibres not be regarded as the result of a process of optimisation? Without touching the questions of the pre-life-to-life transition, criticism of Culberson's argumentation can perhaps be attempted by (a) putting some of the microscopic features of nature's EA into context, together with (b) stressing that optimising genetic material isn't a nonseparable problem and (c) noting how well-behaved most objective functions are on the macroscopic level. On top of that, one could even add that the *truthful friend on a park bench* model is a too nice model for the real-world situation where the noise of luck and circumstance is overlaying all the fitness functions.

Making a lens fit the eye size, making the size of its wings fit the eagle's weight, these are trivial one- or low-dimensional optimisation problems with a fitness maximum in the middle and monotonically decaying fitness outwards. Probably most momentary macroscopic fitness functions acting on populations of living beings can be deemed similarly simple. Just imagine again that all evolutionary steps are gradual, lungs emerged from swim bladders, legs emerged from fishes' fins (that's why some bones in the fins of the coelacanth can be identified with bones in our hands), feathers emerged from reptiles' skin scales. Always there were intermediate purposes. In this respect one can agree with Culberson's notion of "only very localized advantage". These locally trivial optimisation problems on the macroscopic level get translated into harder combinatorial optimisation problems on the level of finding the right modifications of the 21-letter amino acid alphabet coded into the four-letter text of DNA base pairs. But assuming that for proportion changes of tissue structures and organs just a few molecules involved in one or two feedback cycles of gene activity regulation have to be modified very slightly, and moreover assuming that such proportion- and shape-determining regulator genes and molecules exist always with a basic level of diversity in a population, these can be seen as very small-scale combinatorial problems.

[28] Culberson could have also cited Bremermann [55] whose work is discussed in section T.3. He describes stagnation points where two particular mutations of low probability have to come together to lead to further improvement and concludes: "If one would be exclusively interested in biological evolution then one might have concentrated on investigating stagnation phenomena. With an eye on software technology, however, we concentrated on how stagnation could be overcome." The process of biological evolution, obviously, can't tackle this problem by the use of intellect.

Inventing a new type of fibre (e. g. for a spider web) or a new chain of enzymes (e. g. for poison production or digestion), where one or more big proteins have to be re-engineered, translate into much larger combinatorial problems on the genotype level. But why can also these problems be assumed to be smaller than Kauffman's brief calculation suggests? The most important aspect is that also in this case the problem can be seen as separable into smaller subproblems. Take as an example a protein with an enzymatic function that is fixed into the cell membrane by having a hydrophobic base part. Just as the protein has one end with an active pocket responsible for the enzymatic function and one hydrophobic end that likes to be in the cell membrane, the gene code of that protein will be dividable into two main regions as well. And if the code can be divided into two meaningful blocks then it is highly probable that it is indeed divided so in practice by sequences of non-coding DNA. The hydrophobic part can be useful for attaching any other freely moving protein to the membrane, and code for the enzymatically active pocked can be useful if copied and modified for slightly different chemical tasks. This is what "junk code" is for, to be squeezed in between meaningful blocks so the CO operator in the algorithm has a chance of creating useful duplications and dislocations of DNA sequences. The genetic code should be seen as a huge code library of subroutines which can all serve as templates for expanding the library and gaining more functionality for it. Thinking of modifiable templates explains why there are many slightly differing variants of important proteins present in a living being like different types of collagen for use in skin, tendons, cartilage or like myosin for use in different types of muscles. Some forms of these proteins are common to all animals, that means the corresponding code has hardly changed since the separation between animals and other life forms. It is probably not far-fetched to assume that the building blocks necessary for all kinds of gene activity regulation will also come in template-based subroutines. That the systematic division into subroutines separated by junk code reaches down to the level of parts of proteins can be seen by the existence of *exons* and *introns*[29]. The optimisations of different parts of proteins represent separable problems. On top of the separation of gene parts by introns comes the spatial separation of different ends of a protein, which is a chain of amino acids, mirrored in the spatial separation of the corresponding code on the DNA, a pure consequence of the finite size of the polymer molecules.

The decomposition into smaller separable problems is one thing. A second important aspect is that in most cases the exchange of one amino acid will not completely change the way how the whole chain of amino acids is wrapped up into the protein structure. Therefore, even on the level of the combinatorial problem of modifying amino acid sequences, a certain amount of well-behavedness of a protein's performance measure, its fitness function contribution, can be expected. Of course, mutations concerning those few amino acids forming a chemically active pocket will

[29] After a messenger-RNA sequence has been synthesised by transcription of a DNA sequence, some non-coding sections of it, the introns, are cut out and the remaining pieces, the exons, get connected together again to become the mature messenger-RNA. That mRNA is then translated into proteins with the help of ribosomes. See `http://en.wikipedia.org/wiki/RNA_splicing`. The "junk DNA" of introns yields finite ranges along the DNA sequence (in between meaningful gene segments) where CO cuts have an increased probability of not being lethally harmful.

be more sensitive than the ones concerning purely structural regions of the same protein. Modifications of structural backends will make a protein perhaps more or less prone to decay at increased temperature or under acid attack and otherwise impact the geometry or electron density around the enzymatic pocket only a little bit. On other occasions, a point mutation leading to a single wrong amino acid inside the sensitive region of an important enzyme will be lethal for the individual. In that case the fitness function looks pretty simple, too.

To sum up these thoughts: nature tunes the species-dependent evolutionary algorithms herself by inventing concepts like diploidicy and recombination, by introducing even more recombination via chromosomal crossover where also the rate and probable attack points can be tuned by the amount of non-coding DNA in between genes. From simple to complex life forms the genome is a growing library of template-based subroutines. Subroutines get copied, they branch-up, the purposes of the end products shift slowly and incrementally. If a gene collection coding for a lower- or higher-order subroutine is in a stagnation point in Bremermann's sense, then after a shift in purpose this may have changed. Protein building is a discrete combinatorial problem, but it is not like the TSP or graph colouring. On the macroscopic level, one finds the well-behaved fitness functions of engineering problems. These thoughts are proposed as a way around the NFL theorem without being compelled to deny that a dragonfly is a highly optimised system that will teach even some future engineers what they are not yet able to do.

Final remarks on the theory of evolution

The theory of evolution in its fundamental simplicity is extremely successful in terms of being relied on by most people, not only scientists, to explain our experienced world. Few people can outline why quantum mechanics makes more sense than non-quantised mechanics, but most people have a working model of genetic evolution. The theory is part of regular school education, and it is present in televised popular science. However, scientism as an irrational form of extremism has become a real and undeniable issue, not least because the many references to and descriptions of the theory of evolution are so rarely accompanied by general epistemologic thoughts. This is not to be the case here.

Purpose and emergence: Perhaps one should underline that when describing the process of biological evolution playing out within ecosystems, it is often difficult or futile to try to isolate single cause-effect relationships and that the concepts of interdependent development or emergence are more helpful. Plants like raspberries or mountain ashes can be said to have developed a co-evolutionary relationship with birds. The bush invests some of the available resources to produce fruits that can be picked by birds. The service in return is a very far spreading of the seeds which is of crucial importance to pioneer plants like bushes covering the ecological niche of growing in forests in temporary clearings caused by fallen trees or fires. An array of berry-bearing bushes allows several species of small birds to exist year-round in moderate climate zones. Like all co-evolutionary relationships this has nothing to do with purposeful planning, but is merely the result of how the numbers of genome versions in populations play out. Assuming that birds and their behaviour already

existed the last time a plant started buying into the seed spreading service infrastructure, it doesn't seem that wrong to describe this as the plant having developed the fruit with its composition, shape, and colour *for the purpose* of attracting birds.

Here, misconceptions are pre-programmed given the main meaning of the word "purpose" in the context of intelligent planning. Accordingly, should one not be allowed to theorise that cows entered a relationship of co-evolution with humans, exploiting the humans' desire for beef and increasing their own population, capitalising on the humans who help repelling predators like wolves and competitors like gnus, buffaloes, sheep etc.? How should this statement be proven untrue? By numbers cows are currently experiencing a huge success story *because of* their phenotypic features.

Difficulties arise when intermixing our everyday meaning of "purpose" and "means" with a description of the process of evolution, and it gets very problematic where the word "intention" comes in: rabbits do not become faster *in order to* better escape the foxes, the scaly skin of some dinosaur did not develop fringy and hairy structures *in order to* enable Archaeopteryx to fly at some point[30] and hominides did not grow bigger brains *in order to* become better tool users. The theory of evolution does not rely on any outside thinking entity determining intentions.

The application domain of the theory of evolution: The correct spelling is: because of how the numbers and instances of mutations and populations of living beings tend to play out in a setting where statistics counts, we can simply observe the momentary and intermediate results of the evolution process, and developing the theory of evolution upon those observations is in fact very useful for explaining an overwhelming array of facets of nature as it surrounds us. The network of arguments on causalities and effects we call evolution theory indeed turned into the one veritable theory biology has to offer because it enabled us to explain observations that were discovered after the formulation of the theory, e.g. details of molecular biology, or were initially not judged to be in connection with the theory, like the disadvantageous location of the nerve layer in front of the retina in a mammal's eye or the fact that large portions of "junk code" are part of our DNA which becomes only clear after knowing about the function of crossing-over.

That some main concepts of biological evolution can be simulated, i.e. that evolutionary algorithms work, is another nice fact underscoring the status of the modern theory of evolution, or stated alternatively: biological evolution and evolutionary algorithms prove each other.

The explanatory power of the theory of evolution can be helpful to (a) understand observations a posteriori and to (b) predict possible statistical outcomes a priori. Throughout the text there were lots of examples of (a). Examples of (b) could look like the following: if antibiotics are extensively used in animal feeding, this creates the selection pressure background for vast populations of different types of bacteria to optimise DNA code and engineer the right molecules to make them resistant against these antibiotics and to share the newly developed genes among each other.

[30]Rather, it must have *just so happened* that a warming hairy skin structure at some point also began to be beneficial in terms of enabling more and more daring escape jumps down from higher tree branches, ever smoother touch-downs because of pure air friction and only later ever farther jump trajectories turning into glide paths.

Other examples could be predictions in the context of invasive species.

Appeal to nature / is-ought problems (or where not to apply the theory of evolution): Male lions have to wander around, find a pride of other lions and must oust the other male(s) from it in order to find mating partners. Infanticide, the killing of the former male's cubs is often part of the behaviour [339]. From the standpoint of genetic evolution, it is quite clear why it makes sense that the behavioural pattern has emerged. Yet no reasonable person would try to justify the same behaviour among humans by pointing to the laws of evolution. But unfortunately, *Social Darwinism* often comes in more abstract terms, often in less drastic contexts, involving subtler socio-economic issues as e. g. land grabbing or murder. Nevertheless, it is always based on the same fallacy, termed *appeal to nature*: to derive from the observation of *what is* without any further argumentation some normative statements about *what ought to be*.

Is-ought fallacies in connection to notions abstracted from the theory of evolution occur in a wide spectrum of topics. Think of the discussion on the preservation of endangered tropical rain forests. If a particular human fellow argued in favour of the destruction of such a precious ecosystem in order to enable human expansion, adding that the responsibility for any such decision should be passed on to the mechanics of evolution, then I would definitely reject this argumentation as being wrong. But where did the term "precious" suddenly come from and why should "the diversity of the biosphere" become an ethical category? It is true that from the scientific viewpoint we can state that for the populations of a species a certain gene pool diversity is *necessary*, *beneficial*, or better yet *useful* for evolutionary progress. But to deduce from this empirical observation the opinion that biodiversity is *good* in an ethical sense is a transgression of the narrow limits of the scientific viewpoint, and in principle one makes the same mistake as social Darwinism, just in the other direction. If humans ended up drastically reducing the planet's biodiversity and kick evolution into a different trajectory with a temporarily more boring and less aesthetic overall look, the dirtball on which we are living would never care, it has neither brain nor consciousness, and for the time being we must assume the same grade of empathy from the rest of the material universe. But we as humans have self-awareness, can share our experience with it among each other, are able to meaningfully coin terms like "pain", "peace", "cruelty", "mercy", "empathy", we can convene that their meaning becomes evident to everybody through similar experiences and can further convene on the concept of "human dignity". Lastly, the task is up to us to decide or better determine, which way of treatment among ouselves and towards the rest of the biosphere suits our human dignity.

T.3 Pioneering EAs of the early computer age

All genetic information of real life is expressed in the four-letter alphabet of cytosine, guanine, adenine, and thymine[31] and written into DNA and RNA. Computers express all information using the two symbols "0" and "1". Starting out from the consideration of this analogy, the invention of many EA computer codes was the

[31] Uracil takes the place of thymine in RNA.

attempt to simulate or replicate the basics of how biological evolution uses a randomised process to generate meaningful information. The purpose of EAs is just the optimisation of a solution representation. The goals of the pioneering EC work were often wider and included the self-organisation of structure, the emergence of automata or programs, artificial intelligence, artificial life. EAs became popularised by evolution strategies and genetic algorithms. The classic forms of these two EAs are quite narrow in their conceptual scope and they contain just a small selection of the ingredients of MNEA. The high degree of abstraction made comprehensive theoretical analyses possible. The comparison with some of the earlier approaches can show that EAs are always a combination of subroutines or operators from an infinite toolbox of EC concepts which is a more open view than the one of separate schools of ES, EP, GA, GP, etc. expressed throughout much of the corresponding literature of the 1980s and 90s. Some archaeological[32] work on the early EA approaches has been undertaken by David B. Fogel [137–141]. To put modern EAs into context, three of those early works of the 1950s and 60s were selected for brief discussion: research done by Fraser & Barker [145], Bremermann et al. [55], and Reed et al. [382].

1957: Fraser & Barker – Simulation of Genetic Systems

A. S. Fraser & J. S. F. Barker[33] [145] used diploid sets of binary chromosomes with dominant and recessive alleles and implemented gene interactions like *polygenes pleiotropy*, and *epistasis*. Polygene is is the term for several genes determining one phenotypic feature, e.g. a bacterium that can metabolise a certain nutrient only when a combination of several genes comes together which code for different necessary enzymes. Pleiotropy is when one single gene influences multiple phenotypic traits. Epistasis means that the effects of one gene are modified or blocked by one or several other genes. All simulations were done on computers of the ILLIAC family which made the transition from tubes to transistors [215]. The fitness functions in these simulations consisted of the most simple functions able to express (a) the dominant/recessive interaction of homologous genes and (b) the blocking or enforcing interactions between different genes. Optimising the fitness function was only a secondary purpose of the computational experiments. The primary goal was to conduct a simulation of gene pool dynamics. Fraser & Barker wanted to verify or falsify biologists' hypotheses of how evolution works.

Coding and objective function: The key elements of the simulation are as follows. One haploid chromosome consists of n bits which are either 0 or 1. One bit is one gene, so there are two alleles of each gene. If one allele is supposed to lead to a better fitness than the other gene version, and if the different genes are supposed to have not the same effect on the overall fitness, then it is straightforward to formulate the fitness function (of one single chromosome) as a weighted sum $\sum_{i=1}^{n} w_i a_i$ where a_i are the alleles, w_i the weights, and the counter i enumerates the

[32] Interestingly, Bäck, Hammel, & Schwefel also conducted some excavations [20] in which they touched Bremermann's work, but without involving his broad perspective and his lessons learnt into their discussions of the ingredients of ES and GA.

[33] A. S. Fraser is the author of the introductory paper [145] of a whole series of which some papers are authored by J. S. F. Barker.

genes. These contributions of the different genes are so far separable. Next, gene interactions are implemented by introducing multiplicators reducing or increasing the contributions of target genes if, and only if a source gene is present in the form of "1". This explains how to compute the fitness of one chromosome, but how about diploid individuals? If for gene i the maternal chromosome carries the same allele as the paternal chromosome, then there is no problem, they are read out as one single chromosome. For the case of different alleles Fraser defined a dominance factor $h_i \in [0, 1]$, so the primary contribution of a gene if only one allele is 1 is $h_i w_i a_i$ instead of $w_i a_i$. The last ingredient added to come to the final fitness value is a certain amount of noise in order to make the task of the EA more difficult and realistic. The reasoning is to emulate what happens in nature where not only the genotype is a determinant of the effective fitness, since there are also environmental influences determining the abilities an individual can develop, and because also luck and circumstance contribute to its reproductive success. Remember, the goal of the work was not optimisation, but to test under how severe conditions the EA can still lead to solution improvement.

Recombination: In diploid living cells genes are recombined by the random partitioning of chromosome pairs during meiosis and by chromosomal crossover (CO). Thus, any type of gene combination can be achieved, but the chromosomes create *linkage groups* of genes that stay together more often when being passed down to the offspring than pairs of genes from different groups. The gene recombination routine by Fraser is able to divide a linear genome into several linkage groups of genes where the degree of linkage can be chosen arbitrarily.

Selection: The generational cycle follows a simple (μ, λ) scheme, where the population is sorted according to fitness and the first μ elements of the population of λ members will all reproduce with equal probability. Three ways of sorting are initially considered: with descending fitness, with ascending fitness, with ascending distance from the mean fitness.

Fraser's article of 1957 is just the first one of a series extending over several years and reporting on optimisation statistics with different features of the simulation switched on or off, e. g. with or without epistasis, with or without linkage. Fraser saw these Monte Carlo simulations as a method to gain knowledge on the laws ruling the dynamics of "genetic systems" where features like linkage made algebraic solutions to the problem of finding attractors (i. e. stationary states approached asymptotically) impossible [144].

1967: Reed et al. – Simulation of Biological Evolution and Machine Learning

This work by Jon Reed, Robert Toombs, and Nils Aal Barricelli is remarkable because of (a) the evolutionary self-adaptation of strategy parameters determining mutation and recombination behaviour, (b) a co-evolutionary setup, and (c) because of the large examined populations of 10^2 to 10^4 members. The main motivation of the simulations was to check statements made by theoretical biologists on the roles of mutation and recombination for determining gene pool dynamics in competitive ecosystems.

The problem to solve: Individuals have to compete in a simplified game of poker and have to develop optimal betting strategies in competition with each other. Always two individuals compete in 20 rounds of the simple game and the goal is to win more often than the opponent. Individuals are given, by random, one symbol either representing a good or a bad hand of cards at the beginning of each round. The DNA of an individual determines what it does upon receiving the symbol. On good or bad hand it can either bet on high, low, or pass. The costs of the bets "pass", "low", and "high" are 2, 3, and 7 pennies, respectively. Since for each of the two situations "low hand" and "high hand" there are three betting probabilities which must add to 1, there are in total four betting strategy parameters to tune. After both individuals have made their bets, always the one who has placed the higher bet, wins, no matter what the real hands are, and only if the two bets are the same, then the actual hands themselves count and are able to decide the winner or a tie. The game is simple enough, so that game theory can determine the optimal betting probabilities. The co-evolutionary result can then be compared to this benchmark.

Encoding: The chromosomes of the haploid individuals are binary strings concatenating the four betting probabilities and four more variables. The variables are encoded in 3- or 5-bit resolution. The four additional parameters are EA strategy parameters determining mutation and CO behaviour.

Mutation: There are two mutation types. In the first type one of the eight genes is replaced by a random number, in the other one a parameter is incremented by a certain step. The two mutation probabilities and the step size are subject to evolution.

Recombination: The remaining gene determines the probability or ability to reproduce with recombination involving another member of the population (note that the concept happens to be similar to bacterial F-factor genes). Reed et al. experimented with one-point and uniform CO.[34] The one-point CO cuts anywhere through the bitstring. The uniform CO redistributes genes, not bits. Fogel points out that the GA community rediscovered uniform CO only 22 years later [137]; from the view of optimising engineering problems it is clear a priori that the locations of and distances among parameters on the chromosome should generally rather not affect the optimisation process.

The algorithm: The population is divided up into pairs of individuals playing 20 rounds of the poker game against each other. All losers are erased from the memory, only the winners remain. Individuals with the right setting of the corresponding gene are assigned to other individuals and the pairs undergo recombination. The empty places are filled up by copies which undergo mutation.

1968: H. Bremermann et al. – Numerical Optimization Methods Derived from Biological Evolution Processes

The work of Hans Bremermann and his team is often forgotten[35] when describing the history of genetic algorithms. They were not only the first to make the set of abstractions to create a standard GA [138], they were also the first to have put

[34] These CO operators are explained in section T.4.4 on page 527.

[35] e. g. in [104], whereas references to their work are present in [173]

the standard GA aside for the consideration of real-world problems in continuous optimisation [55]. On their inventions[36] one can remark that they generalised and explored different versions of EC concepts where others stuck to narrow scenarios for decades. On their rejection of the standard GA, one can refer to the very simple argumentation[37] found in [55] and represented in figure T.11.

In the conference article on "Numerical Optimization Methods Derived from Biological Evolution Processes" [55] Hans Bremermann recollects lessons learnt during several years of research on optimisation procedures involving the laws of genetic systems and other stochastic processes. For him, his work is inspired by the archetype of biological evolution, but targeted at software[38] development. This sets him apart from the other two mentioned examples of pioneer EAs. He studied fitness functions with one single valley and monotonically ascending smooth slopes. In [55] only the case of real-coded search spaces, $\vec{x} \in \mathbb{R}^n$, is dealt with. The class of functions to be optimised can be described as norms in a transformed space: $f(\vec{x}) = ||A\vec{x}-\vec{b}||$, where A is a non-singular square matrix. If $A = I$ then f is a spherical cone or funnel. Generally, A stretches, squeezes, and rotates the funnel so it can yield landscapes looking like the one in figure T.11. Arbitrarily transformed funnels are seen as a test problem representative of the close vicinity of local minima in any real-world problem where the fitness function is sufficiently smooth. Think of linear feedbacks and quadratic potentials for approximating stable equilibria in physics. (One can add that if selection is based on rank instead of fitness, it doesn't matter whether the slopes mount linearily or with the n^{th} power of distance from the minimum, as long as the isocontours look the same.) The demand is that any algorithm should at least be able to solve this case before moving on to more complicated problems. But simple real-coded GAs fail because in a setup where vector components x_i are mutated independently with relatively low probabilities (so that on most occasions only a fraction of a vector's components change) the narrow corners of the isocontours represent stagnation points from where there is no way out. An explicative diagram is shown in figure T.11. Bremermann reconstructs a rich learning process of his team experimenting with different algorithm tweaks to get around this problem. One early step was to introduce a recombination operator based on differences (as in differential evolution, invented in 1994 [440]). Another step consisted in an algorithm working purely on mutation and activating the recombination operator only after stagnation detection, resulting in a landscape-responsive metaheuristic. A next step was line search along vectors determined by a subroutine approximating the local gradient, thus a completely deterministic search procedure. Finally he

[36] Fogel remarks in [138]: 'Indeed, by 1962 there was nothing in Bremermann's algorithm that would distinguish it from what later became known as "genetic algorithms" based on research at the University of Michigan [...]. Furthermore, Bremermann's evolutionary algorithm already included real-valued encodings, in addition to binary encodings, thereby anticipating the move to "real-coded genetic algorithms" that occurred 30 years later'. Another innovative concept is also noteworthy: in Bremermann's article [55] (p. 613) there can also be found a geometrically motivated heuristic where a step size parameter λ is steadily reduced in a feedback loop. The argumentation builds on estimating what fraction of a sphere that can be reached by a fixed-distance translation step lies within a conical volume where improvement can be expected. This yields an interesting comparison of concepts in relation to step size adaptation schemes in ES literature.

[37] to be compared with a sample of an exerted explanation of 1995 [394] of a perceived novelty

[38] In 1968, he still felt the need to put the word "software" within quotes.

describes how getting rid of the sophistication of the gradient examination routine made the algorithm more efficient: just repeating simple 1D minimisations along isotropically distributed random directions led to faster optimisation because of the saved cost. This may come as a counter-intuitive conclusion, but it is in line with the main ideas behind classic evolution strategies. If at every moment the problem is only one-dimensional and the search follows the landscape's gradient only on the average, this is a similar way to prevent the explosion of the computational cost with increased dimensionality as in simple ESs. The advantage is that steady improvements are made right from the start of an optimisation run while knowledge on the local gradient comes in and is updated piece by piece.

Bremermann's article [55] conveys an open-minded spirit of problem-oriented algorithm engineering. It says that seeking inspiration in nature is good, as long as one keeps in mind that nature underlies her own restrictions and that for an intelligent programmer who has other restrictions, nature's algorithmic ingredients need not be the best and only choices there are.

This author values the texts by Fraser, Reed, and Bremermann [55, 145, 382] as precious reading experiences with the power to prevent a contraction of the thinking horizon when one is beginning to dive into EA literature and when this danger arises depending on the choice of the first books and articles read and websites visited.

T.4 Overview of widely used evolutionary algorithms

This author shares the opinion that a too deep distinction of the different and supposedly separate schools of evolutionary computation prevents the full exploitation of the complete ensemble of useful possibilities for the design of EC methods. Instead of carrying on a separation of different historical lineages of ideas, the field of EC should rather be seen as an ensemble of concepts and ideas offering a diverse toolbox for the curious user to compose algorithms fitting particular applications. Therefore, the following sections, where popular EAs are sometimes described under the names of the historical branches, should not be understood as reflecting separate subdivisions of EA, but rather as a short list of instructive algorithms examples. The explained list of algorithm examples is thought to offer easier understanding of the impact of certain conceptual aspects on the character of the resulting search strategies than a mere enumeration of the tools in the box could. The list enumerates important historical approaches (ES and GA) on the one hand, and some modern EAs on the other. The first set shows how much freedom there is for making particular abstractions of natural evolution focusing on different aspects of it. The other part, however, yields the insight that greater efficiency is often being achieved by incorporating ideas alien to natural evolution. A reader with a pure interest in efficiently applying a state-of-the-art EA to a new problem, might consider focusing on the modern approaches among the following paragraphs and expand his reading to (a) those web resources, where good overviews as well as instructive code projects can be found: [1, 61, 191, 279, 337, 373, 403, 430, 439, 513, 514], to (b) recent overview books and review articles in dedicated journals (at this moment e.g. [62, 101, 114, 280, 389]), and to (c) the published results of competitions on EAO of which the two most important are the *Black-Box Optimization Benchmarking (BBOB)* [192] com-

petition held at the *Genetic and Evolutionary Computation Conference (GECCO)* and the competitions on real-parameter single-objective optimisation [443] held at the *IEEE Congress on Evolutionary Computation (CEC)*.

T.4.1 Evolution strategies (ES)

A first basic form of the algorithm was proven to work in the early 1960s by a student named Ingo Rechenberg without the use of any automated computer [537]. Instead, there was an experimental setup with a set of n tunable parameters and one observable $\in \mathbb{R}$ classifying the outcome and serving as objective function. The simple pseudocode is shown in algorithm 4.

Algorithm 4: Ingo Rechenberg's first ES

```
while g <= G and not StopCriterion() do
    decide by random which parameter x_i to change next
    determine another random number δ (the amount of change)
    carry out that change x_i ← x_i + δ in the experimental setup
    measure the observable
    if fitness improved then keep the change
    else restore old setting
```

In the μ-λ-notation invented later by Rechenberg this simple hill-climbing scheme corresponds to a $(1+1)$-ES. The main driver of ESs is the mutation operator, which changes a chromosome $\vec{x}$ by adding a small vector $\sigma\vec{\xi}$ pointing into a random direction and scaled by a scalar mutation step size parameter σ. Implementing an isotropic distribution of mutation step directions ensures that the search is independent of the coordinate system. Let us look at several basic examples of evolution strategies expressed in the μ-λ-notation. In a $(1+5)$-ES five offspring points are generated from one parent vector by use of the mutation operator. That vector among all six points having the best objective function value will be the parent vector for the next generation. In a $(1\,,5)$-ES the offspring generation process is the same, but the parent vector is excluded from the selection of the next new parent. In a $(\mu+\lambda)$-ES the best solution is always conserved, whereas in a $(\mu\,,\lambda)$-ES this is not the case. An upper index G can indicate the maximum number of generations in the notation, e. g. $(1\,,5)^{G=20}$-ES. If there are several parents, like in a $(3+5)$-ES or $(3\,,5)$-ES, then there are two straightforward ways to proceed. The first option is to produce each new offspring vector by randomly choosing one of the three parents for copying and then apply mutation. The second option, which is more often present in classic ES literature, consists in computing the mean of the three parent vectors and using that point as base from where to do the random mutation steps. The notations for that second version are $(\mu/\mu+\lambda)$-ES and $(\mu/\mu\,,\lambda)$-ES. This kind of averaging over successful trial vectors is the only kind of recombination operator used in evolution strategies.

It has to be noted that evolution strategies do not care about gene pool diversity. Any $(1\dagger\lambda)$-ES or $(\mu/\mu\dagger\lambda)$-ES can be seen as a population cloud anchored at one

specific point in the search space. The question is just how that point moves over time. If $\mu > 1$ and no averaging is used, then different corners of the search space can theoretically be searched in parallel. But in reality, the weaker lineages will always go soon extinct. The result is that a simple ES is inherently set up to be a very local search engine. The search behaviour can be described as the movement of an amoeba, because a round population cloud reaches out into space and iteratively re-anchors itself where better fitness values are found. The mutation step size parameter $\sigma \in \mathbb{R}^+$ is the most important control parameter because it determines the size of the population cloud and how fast it can move in the search space. Thus it is also the quantity deciding on the balance between exploration (sparsely searching vast areas when σ is large) and intensification (denser probing when σ gets smaller). Here are some basic ways of implementing a mutation operator generating a new point $\vec{x}' = \vec{x} + \sigma\vec{\xi}$

1. sampling the mutation step from an n-dimensional normal distribution $\sigma\vec{\xi} \in \sigma\mathcal{N}(\vec{0}, I)$, where $\mathcal{N}$ designates the normal distribution and I is the identity matrix,

2. the mutation steps go into random directions but are all of the same distance, i. e. $\sigma\vec{\xi} = \sigma\vec{\xi'}/|\vec{\xi'}|$ with $\vec{\xi'} \in \mathcal{N}(\vec{0}, I)$,

3. the same as above with a uniform distance distribution, i. e. $r\sigma\vec{\xi} = r\sigma\vec{\xi'}/|\vec{\xi'}|$ with $\vec{\xi'} \in \mathcal{N}(\vec{0}, I)$, and $r \in [0, 1]$ to be generated from a uniform distribution $\mathcal{U}(0, 1)$,

4. with individual step size parameter for each direction in the coordinate system, then $\vec{x}' = \vec{x} + \vec{\xi}$ with $\vec{\xi} \in D\mathcal{N}(\vec{0}, I)$, where the diagonal matrix $D = \mathrm{diag}(\sigma_1, ..., \sigma_n)$ stretches the distribution more or less along the different coordinate axes (this is the only coordinate system-dependent implementation in this list),

5. sampling the mutation step from a multivariate normal distribution which can be arbitrarily rotated, i. e. $\sigma\vec{\xi} \in \sigma\mathcal{N}(\vec{0}, C) = \sigma C^{\frac{1}{2}}\mathcal{N}(\vec{0}, I)$, where the matrix $C^{\frac{1}{2}}$ stretches and rotates the normal distribution $\mathcal{N}$ (if C is the covariance matrix of the recently successful mutation steps, then it is a CMA-ES, see section T.4.6).

The method used for generating the mutation steps determines how the population cloud looks, and σ is the tuning parameter scaling the cloud's size.

Making σ responsive to the scanned landscape has been a concern since the early beginning of ES. The second method of the list, where the steps are in random directions but with a fixed length, offers a very simple possibility for step size adaptation. One only needs two different step sizes $\sigma_1 < \sigma_2$ which are applied for offspring generation with probabilities $P_1 = P_2 = 0.5$. Every couple of generations it can be looked back and analysed which one of the two mutation patterns has contributed more often to fitness increases. If it was the far jumps then increase σ by multiplication with a step size adaptation factor $s > 1$, and if it was the short jumps then decrease it by division by the same factor. In the μ-λ-notation the step size adaptation factor can be indicated by a subscript, e. g. $(4\,,12)^{G=40}_{s=1.2}$-ES.

A not very different step size adaptation scheme can be implemented based on a supra-ES scheme on the level of populations. From the current best solution two little ES with a fixed number G of generations can be started. The two ES are equivalent except that they use constant but different step sizes. If the one with the larger steps turns out to win the competition, then increase the step size base value, otherwise decrease it, so the new step size is either $\sigma_{\text{winner}} \cdot s$ or σ_{winner}/s. Then the next competition starts. In the generalised μ-λ-notation where edgy brackets designate the level of populations that scheme can be expressed as $[\mu' \overset{+}{,} \lambda'(\mu \overset{+}{,} \lambda)_s^G]_{s'}^{G'}$-ES where $\mu' = 1$ and $\lambda' = 2$. As the two competing ES may sometimes be attracted by different local optima, that scheme can also be seen as borrowing from MNEA the concept of population isolation with the intention to increase gene pool diversity and the robustness against local optima.

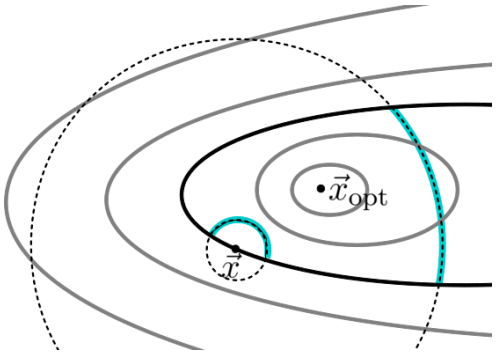


Figure T.8 Explaining the $1/5^{\text{th}}$-rule for step size adaptation.
Point $\vec{x}$ designates a known solution candidate, point $\vec{x}_{\text{opt}}$ the minimum to be found, the elliptic lines show isocontours of the objective function. The two dashed circle lines connect all points that can be reached by mutations with a fixed mutation step size σ; there is one small circle that can be reached after mutations with σ_1 and a larger circle connecting points reachable by bigger steps with σ_2. Both circle lines possess one segment, shaded in the sketch, designating the subsets of those mutations leading to improvement; for the smaller circle this subset covers almost one half of the circle line, which can be attributed to the general fact that in landscapes with a minimal degree of smoothness a small enough σ can always be found so that the circle of mutation explores only the local gradient and the fraction of improvements tends towards 50 %. For the larger circle, in contrast, it covers a substantially smaller fraction. Based on those considerations, the simple but efficient rule was devised to increase the step size parameter after one or a few generations of more than $1/5^{\text{th}}$ of the mutations leading to improvement, and to decrease it otherwise. This leads to the following behaviour: the population cloud wandering through the search space expands and speeds up when it follows a gradient and it slows down and zooms in near a minimum. From the sketch it can also be inferred that approaching the optimum point $\vec{x}_{\text{opt}}$ on a horizontal line from left or right will take substantially more generations than coming in from above or below.

There are many more ways to make σ a state variable of the algorithm, but one of the simplest and most widely used rules for classic ES is the $1/5^{\text{th}}$-rule [379] (see fig. T.8). Here, the step size σ is controlled by a feedback loop where the input is the percentage of mutations that have led to fitness improvements measured on a batch of steps made in the recent past. From the case shown in figure T.8 it can be shown that in smooth landscapes this feedback loop will always tend to settle on a finite equilibrium value of σ. But it can also easily be deduced that the $1/5^{\text{th}}$-rule will fail in following down a valley if its cross section is V-shaped instead of U-shaped so the isocontours form sharp angles. When the angles are smaller than $360°/5 = 72°$ then the step size of a population caught on the edge will quickly go towards zero. No matter how small the mutation steps, the success rate will be the same.

There is a particularly interesting aspect to the $1/5^{\text{th}}$-rule in that it is a conceptual step away from the workings of evolution in nature and towards a swarm algorithm. Information is not any longer only processed longitudinally along lineages of descendants and leading to different results at the ends of different branches. Instead, information about the success statistics of the whole population is shared as it comes in and the whole population enjoys the more quickly adapted mutation step size.

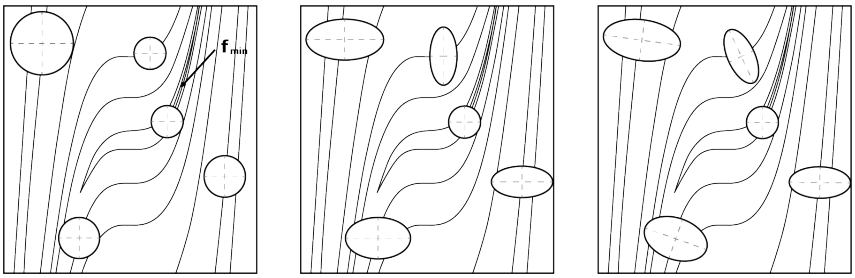


Figure T.9 Probability isocontours depending on mutation distribution schemes. The picture on the left shows examples of isocontours if one single mutation step size applies to all dimensions. The middle image shows the situation if different mutation step size parameters σ_i are used for each parameter x_i. This leads to ellipsoids which are always aligned with the coordinate axes. The plot on the right, however, shows the ultimately desired distribution isocontours for search agents (individuals or populations) located at different positions in the search space if the agents are to perform the local part of the search with an optimal progress rate. For this to happen, an agent needs an idea of the local gradient vector and a multivariate distribution of mutation steps has to be rotated into the same direction. (Sketch layout adapted from Bäck et al. [21].)

Thinking of a generic parameter-tuning problem where the optimised system reacts with higher sensitivity to changed settings of some parameters and is less sensitive to others, it can easily be motivated to go from one general step size to n independent parameters σ_i. A simple adaptation scheme for the data in the vector $\vec{\sigma} = \{\sigma_1, ..., \sigma_n\}$ can be implemented most naturally by giving each individual its own vector $\vec{\sigma}$ and make it also subject to the same conditions of mutation and inheritance as the proper chromosome [21, 379, 407]. The result will be populations with mutation probability density isocontours like the ones shown in the middle of figure T.9. That diagram shows that the scheme may be efficient if the motivating assumptions are valid, but it can't produce ideal distributions in arbitrary fitness landscapes. The desire to create the ideal distributions shown in the right diagram of T.9 leads to the development of the CMA-ES, the currently most efficient and popular ES, that will be described below in section T.4.6.

In contrast to many other EAs, ESs have shown to offer feasible ways to analytically estimate [21, 41] progress rates of the populations in certain well-defined search landscapes. Important results are scaling laws of the computation cost with problem dimension. Despite the method's simplicity and the seemingly heavy influence of the random number generator, simple classic ESs can be competitive local search engines because the computation cost scales with $\sqrt{n}$ and not with n as would be the case using a deterministic method determining local gradients from probing $n + 1$ closely neighbouring points [380]. However, even with temporarily isolated parallel populations, ESs tend to search rather locally. Stepping up λ, λ' and γ' quickly leads to very high numbers of calls to the evaluation function with

limited gain in performance of global search in difficult landscapes. Random restarts should be considered instead of making one single trajectory ever more costly when examining multimodal landscapes. Since the progress of a population relies on it covering a small enough subspace to be pressured by a local fitness gradient, and since the mutation operator is always tuned to reduce the population's covered subspace to such a small size, one such population often tends to be not much more than one single (approximative) gradient probe examining one point in the search space. Such basic forms of ESs can only be usefully applied to a small amount of very well-behaved engineering problems, where one expects a monotonical descent to the global minimum.

T.4.2 Simulated annealing (SA)

Simulated annealing, proposed by S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi [234] and based on the Metropolis algorithm [297], can be seen as a $(1+1)$-EA with a special offspring acceptance criterion. There is one parent generating one offspring at a time through a mutation operator generating small moves. In case the mutated solution is accepted, it replaces the parent. Improvements are always accepted and mutants deteriorating the fitness are accepted with a probability depending on two circumstances, namely the quantity of deterioration and an external control parameter T. The control parameter T (named T for temperature) is used to make the algorithm less and less likely to accept deteriorating moves during one search, that means the search is tuned more and more greedy. Often it makes sense to use T to control also the size of the mutation steps. This is always the case when larger steps can be expected to lead to larger fitness variations on average, since otherwise the algorithm would not have any optimisation power in the initial phase because of accepting almost every move, and it would lose its character and turn into pure greedy search (only accepting improvements) at a later stage. SA can be seen as a single particle bouncing around in the search landscape and slowly losing its kinetic energy so the upwards jumps become ever rarer and smaller, the search becomes more localised and the particle ends up bouncing down into one local valley. The algorithm looks like this:

SA is also a nature-inspired algorithm, but the inspiration comes from physics instead of biology. Annealing is the process of hardening a freshly manufactured metal piece by heating it up to glowing temperature and letting it cool down slowly again. What happens is that the energy input leads to an increase of the atoms' brownian movements. The heat lets the atoms bounce between alternative places in the atomic grids of the polycrystalline structure. The temperature decay lets them slowly settle at the locations where they have the lowest potential energy in their neighbours' potential wells. When the material cools and the energy is taken away, it means that dislocations and grain boundaries are moved, diminished, or removed, the atoms and neighbour distances are redistributed more evenly, and this results in the metal part becoming able to bear higher structural loads and offering less crack nuclei. After annealing the atomic structure is in a state of lower potential energy, so the overall process is exothermal, but the heat is necessary to overcome the many little energetic barriers on the way between the initial and the optimised

Algorithm 5: The simulated annealing algorithm

```
g ← 0
T ← T_0
x⃗ ← RandomPoint()
f ← evaluate(x⃗)
while g <= G and not StopCriterion() do
    x⃗′ ← mutate(x⃗,T)
    f′ ← evaluate(x⃗′)
    Δf ← f′ − f
    if Δf < 0 then
        x⃗ ← x⃗′
        f ← f′
    else
        P ← exp(−Δf/T)
        if rand() < P then
            x⃗ ← x⃗′
            f ← f′
    T ← CoolDownRule(g,T)
```

configuration.

If the SA algorithm is applied to optimisation problems reflecting the motivation case, like finding particle distributions of minimal potential energy E, then it functions really as a physics simulation engine and the greediness control parameter T has the meaning of temperature. The salt of the algorithm lies in the acceptance criterion which tolerates a worsening of E with a probability P determined by a Boltzmann factor $P = \exp(-\Delta E/T)$ because the repetitive application of this kernel of the Metropolis algorithm leads the system into states which are in thermodynamic equilibrium at the temperature T. If the temperature decays slowly enough so that the system is kept through many mutations close to the thermodynamic equilibrium, then it becomes extremely improbable or practically impossible for the system to get stuck in tiny local ditches high up in the energy landscape.[39] Another advantage of SA is that other concepts of statistical physics and thermodynamics can be carried over. An example is measuring the heat capacity $C(T)$ of the system being annealed and optimised, i. e. the energy lost per temperature increment, to diagnose the status of the search. In large systems peaks in $C(T)$ indicate phase changes (changes in the way or degree of ordering). If the heat capacity goes to zero, then the system must be frozen.

But how about optimisation problems which do not reflect the motivation case?

[39] In literature (also in [234]) it is often stressed that the tolerance towards deteriorating steps is the crucial ingredient of the algorithm allowing to overcome "energy barriers" in the search landscape. But it should be added that in a normal kind of search algorithm the mutation operator ignores intermediate states, i. e. mutation is something like *tunnelling*, so the mutation operator itself enables jumping over energy barriers as long as it executes finite step sizes, independently of any acceptance criterion.

Problems where the thermodynamic analogies are not applying? Kirkpatrick [234] already notes that SA can be deemed useful in cases where there are "many nearly degenerate random ground states rather than a single ground state with a high degree of symmetry" and "that the analogy between cooling a fluid and optimization may fail" in cases where a "typical optimization problem will contain many distinct, noninterchangeable elements, so a regular solution is unlikely." If we wanted to translate this into descriptions of objective function landscapes, it means that SA is suitable for landscapes where the space is scattered with many valleys which all have nearly the same function value at the bottom, where it does not really matter by which of the many attractors the search agent is being caught, and where only a series of small energy barriers in the form of little wrinkles have to be overcome to follow the underlying gradient leading towards the attractor. After all, SA, like any other steadily converging $(1 + 1)$-EAs ends up examining only one single valley more closely than all the others and must therefore be called a strictly local search method.

The modification of the method by Corona et al. [91] alleviates the disadvantage of sticking to the trajectory of one single particle by allowing a basic form of history branching. In their implementation the temperature does not decay gradually, but is multiplied with a reduction factor every N_T iterations of the algorithm kernel. The next N_T iterations do not continue the currently drawn trajectory, but instead start out at the best location encountered during the latest temperature step or the best found so far. The thought may have already been seeded in [234] because Kirkpatrick speaks of the procedure generating "a population of configurations" while evolving the system at a given temperature. An obvious modification to SA is to go to real populations of particles in the state space and to introduce recombination operators mixing chromosomes. Web searches in that direction yield plenty of articles.

The case of simulated annealing is quite interesting for learning about nature-inspired methods. It is again one example where conceptually moving beyond the pure mimicking of natural phenomena, i. e. not letting "nature-inspiration" become a dogmatic prison, helped to create more efficient algorithms. On the other hand, the fact that the $(1 + 1)$-EA version of SA is not suited to searching objective functions where the final choice among many attractors matters, reminds us not to forget about where the method came from. Having the existing consequences of the nature-inspired background in mind can help to recognise more easily those cases where problem and optimiser do not fit together.

T.4.3 Genetic algorithms (GA)

Classic genetic algorithms, very much like evolution strategies, designate one particular, strong abstraction of natural evolution. Whereas ESs emphasise incremental mutations of real-coded chromosomes, classic GAs concentrate on binary-coded chromosomes being mainly the subject of a set of recombination operators. When the term 'genetic algorithm' was coined during the 1970s through works of John Henry Holland, including his book "Adaptation in Natural and Artificial Systems" [209], and co-workers from his department, e.g. Kenneth Alan De Jong [106], the algorithm itself was nothing really new, according to Fogel [138, 139]. GAs became the

most popular population-based EAs of the 1980s and 90. The GA in its most simple form is outlined below in algorithm 6. (For parent and offspring generations the shorthands F0 and F1 are borrowed from biology.)

Algorithm 6: basic GA

$g \leftarrow 0$
$\mathcal{P}_{\mathrm{F0}} = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_N\} \leftarrow$ `GenerateRandomPopulation(`N`)` `// parents`
$\mathcal{P}_{\mathrm{F1}} = \{\vec{x}'_1, \vec{x}'_2, ..., \vec{x}'_N\} \leftarrow$ `GenerateEmptyPopulation(`N`)` `// offspring`
`EvaluateMembersOf(`$\mathcal{P}_{\mathrm{F0}}$`)`
while $g <= G$ *and not* `StopCriterion()` **do**
  for $j \leftarrow 1$ **to** N **do**
    $k, l \leftarrow$ `SelectionRules(`$\mathcal{P}_{\mathrm{F0}}$`)`
    $\vec{x}'_j \leftarrow$ `CrossingOver(`$\vec{x}_k, \vec{x}_l$`)`
    $\vec{x}'_j \leftarrow$ `Mutate(`$\vec{x}'_j$`)`
  `EvaluateMembersOf(`$\mathcal{P}_{\mathrm{F1}}$`)`
  $\mathcal{P}_{\mathrm{F0}} \leftarrow \mathcal{P}_{\mathrm{F1}}$ `// offspring become new parents`

There can be many modifications of this basic template. One example has to do with whether to use one or both chromosomes resulting from a recombination operation involving two parents. In algorithm 6 just one resulting chromosome is copied into the next generation, whereas the operation yields two offspring chromosomes in algorithm 7. In the latter case each selected parent will get its entire chromosome copied into the next generation, it will be present in small parts in different individuals, and only small further bit-wise deviations will be introduced by mutation. In the first case, more parents get selected, but for some this means that only a small part of the chromosome will be present in the next generation.

Algorithm 7: basic GA (alternative version)

$g \leftarrow 0$
$\mathcal{P}_{\mathrm{F0}} = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_N\} \leftarrow$ `GenerateRandomPopulation(`N`)` `// parents`
$\mathcal{P}_{\mathrm{F1}} = \{\vec{x}'_1, \vec{x}'_2, ..., \vec{x}'_N\} \leftarrow$ `GenerateEmptyPopulation(`N`)` `// offspring`
`EvaluateMembersOf(`$\mathcal{P}_{\mathrm{F0}}$`)`
while $g <= G$ *and not* `StopCriterion()` **do**
  $\mathcal{P}_{\mathrm{F1}} \leftarrow$ `SelectionRules(`$\mathcal{P}_{\mathrm{F0}}$`)`
  $\mathcal{P}_{\mathrm{F1}} \leftarrow$ `CrossingOverAmongMembers(`$\mathcal{P}_{\mathrm{F1}}$`)`
  $\mathcal{P}_{\mathrm{F1}} \leftarrow$ `MutateMembers(`$\mathcal{P}_{\mathrm{F1}}$`)`
  `EvaluateMembersOf(`$\mathcal{P}_{\mathrm{F1}}$`)`
  $\mathcal{P}_{\mathrm{F0}} \leftarrow \mathcal{P}_{\mathrm{F1}}$ `// offspring become new parents`

Algorithms 6 and 7 are (μ, μ)-EAs where the whole population is replaced in each generation and the best solution is not conserved. A steady-state GA [450, 509] is another common version. Here, the best solution is being conserved because only the worst current chromosome gets replaced in each iteration. So the steady-state GA is a $(\mu + \lambda)$-EA with $\lambda = \mu + 1$ and it features a more aggressive search

characteristic [507]. The outline of this procedure is shown in algorithm 8.

Algorithm 8: basic steady-state GA

$\mathcal{P} = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_N\} \leftarrow$ `GenerateRandomPopulation(`N`)`
`EvaluateMembersOf(`$\mathcal{P}$`)`
while *not* `StopCriterion()` **do**
  $j \leftarrow$ `FindWorst(`$\mathcal{P}_{\mathrm{F0}}$`)`
  $k, l \leftarrow$ `SelectionRules(`$\mathcal{P}_{\mathrm{F0}}$`)`
  $\vec{x}_j \leftarrow$ `CrossingOver(`$\vec{x}_k, \vec{x}_l$`)`
  $\vec{x}_j \leftarrow$ `Mutate(`$\vec{x}_j$`)`
  `Evaluate(`$\vec{x}_j$`)`

In these descriptions, the metaheuristic level of the GAs is given, where the genotype representation and the details of the mutation and recombination operators remain unspecified. In a canonical GA, which is the classic or standard case [507], (a) the genotype representation is a binary string, (b) the parent selection routine ensures reproductive success proportional to fitness, and (c) one-point and two-point crossing-over are used as recombination operators. Figure T.10 shows the effects of the three traditional operators of canonical GAs (mutation of a bit, one-point CO, and two-point CO) on a binary chromosome.

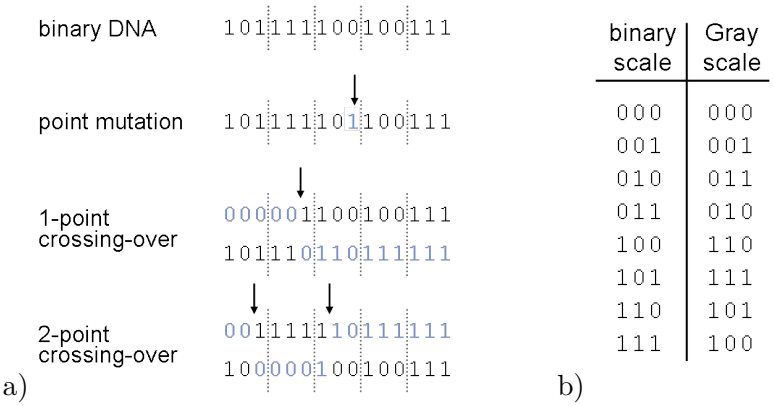


Figure T.10 Binary chromosome.
a) In a binary-coded genotype representation the parameters x_i of a parameter-tuning problem are expressed as binary numbers. In this example the chromosome is a binary string of a length of 15 bits, created by the concatenation of five parameters represented in three-bit resolution. The numbers in blue show modifications through mutation (bit flip) and recombination operators. In a binary-coded chromosome any data modification has a substantially greater impact if it concerns the first few digits of an encoded solution parameter, whereas modifications of the last few digits are marginal in a well-resolved encoding. If a phenotypic feature has to grow incrementally following a gene drift, then the genotype representation poses different hurdles: from 011 to 100 all bits have to be modified whereas from 100 to 101 just one bit flip is necessary. If this behaviour is to be avoided, then the Gray scale encoding can be utilised. The main characteristic of the Gray scale is that for each step change of the discretised parameter only one bit has to be changed. Since the Gray scale is also cyclic, only one bit flip separates the last from the first number in the listing. But still the last few bits of a parameter have marginal impact compared with the first few ones.

However, if one is targeting engineering optimisation problems, where n real-valued parameters $x_i \in [a_i, b_i]$ have to be tuned, why should one go through the effort of devising the binary genotype representation and defining the binary operators?

Table T.1 Why genotype representation matters in a GA.
Applying the metaheuristic of a canonical GA to a continuous domain optimisation problem and switching between binary- and real-coded chromosomes changes the behaviour of the search substantially. The reason is that the low-level heuristics are not the same in the two representations. This table lists important details from where the differences arise. Only the basic operators common to canonical GAs are considered here: one-point CO, two-point CO, and point mutation. Point mutation is an operator acting on bits or genes x_i, not on whole chromosomes $\vec{x}_j$. The CO operators cut between bits or genes and swap parts of chromosomes. (No mutation or recombination operators based on geometric operations in the search space are considered.)

feature	binary-coded GA	real-coded GA
mutation	bit flip (but implementations are possible where genes are randomised)	addition of a number $\delta \in \mathbb{R}$ to one chromosome component x_i; the distribution generating the random numbers has to be carefully chosen
mutation rate	generally low; purpose is to reintroduce the missing allele if for one locus i: $x_i = y_i \ \forall \vec{x}, \vec{y} \in \mathcal{P}$	low mutation rate can lead to stagnation as sketched in fig. T.11 when mutation combinations are needed but don't happen
modification locus	using binary code the impact of modifications is different depending on where the modification happens, flipping the first digit of a gene has a huge impact, flipping the last one has marginal impact	no influence per se
CO cutting points	between any pair of bits (but special implementations are certainly possible where cuts are allowed only between genes)	between any pair of genes, i. e. vector components
gene linkage	closely neighbouring genes get separated by CO less frequently than distant genes; in case of one-point CO it matters whether genes lie near the centre of the chromosome or near the ends	
CO generates new information	yes, if a cut lies inside a gene; if gene i is represented in one parent as 1111 and in the other one as 0000 then there are three different places to cut, all generating a new pair of maybe yet unseen genes; two consecutive COs with cutting point shifted by one digit are equivalent to a bit flip	no (therefore the mutation rate must be higher as in a binary-coded GA if gene pool diversity is not to be lost quickly)

Why not simply apply the GA metaheuristic to the list $(x_1, ..., x_n) = \vec{x}$ of parameters in $\mathbb{R}$? The answer is that the genotype representation really matters, that the same MH will hardly lead to the same search algorithm. The representation question is not just a matter of implementation convenience. The two ways of implementation will lead to two very different search algorithms acting in two very different search space topologies. Table T.1 lists the features from where the differences arise.

Having the example of MNEA in mind, one would normally say that random mutations generate new genetic information and recombination makes it possible that all the new data structures can be tried out in many combinations with other genes. The effects highlighted in table T.1 make it clear that in a real-coded (RC) genetic algorithm this view is correct. But in a binary-coded (BC) GA it is very different. Here, the CO operations contribute greatly to the generation of new information. Remember, no new individual is created without CO in a GA. It means

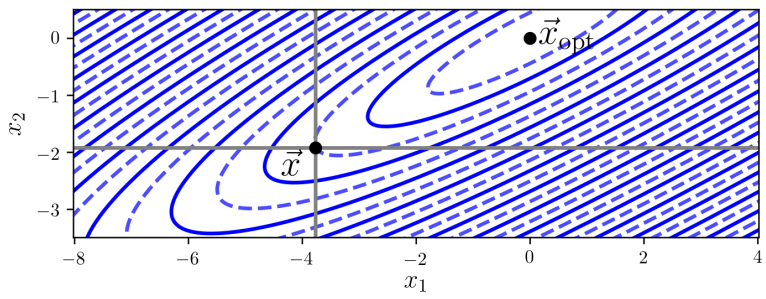


Figure T.11 GA Stalling on diagonal valley.
An EA relying on a low probability of component-wise mutation of a population $\mathcal{P}$ of chromosomes $\vec{x}_j$ will stall on narrow valleys which are not aligned with any coordinate axis. Such diagonally oriented valleys can e. g. be generated by transforming spherical valleys with a matrix of high condition number. The interesting question is: what type of mutation operation can lead to an improvement starting from point $\vec{x}$? By mutation of a single vector component only the points forming the crosshair can be reached. Only very few of these points mean improvement, and they can be only found among the smallest steps. If the valley had sharp edges and the isocontours the shapes of diamonds instead of ellipses, then no improvements were possible at all after a point $\vec{x}$ has been caught on the ridgeline connecting the sharp diamond tips. For a component-wise mutation operator to lead to improvement in such a case, two mutations into the right directions have to come together. The higher the condition number of the valley, the smaller the fraction of beneficial mutation combinations. If such combinations correspond to the multiplication of small probabilities in a given EA, then this EA can easily stall in these scenarios.

that the CO operators together with the selection pressure make up the main search engine, and the mutation operator has only the secondary function of oiling that engine by constantly supplying a minimal gene pool diversity at each bit position. But carrying a low mutation rate which works in a BCGA over to a RCGA creates the problem of stagnation, explained in figure T.11.

How is the selection pressure implemented in a traditional GA? The answer is fitness-proportional selection. Take the example of a parent population $\mathcal{P}_{F0}$ with three members and fitness values (to be maximised) of 5, 3, and 2. Then, when generating the three members of $\mathcal{P}_{F1}$, each time a parent needs to be selected, the probabilities for the three members of $\mathcal{P}_{F0}$ to be chosen would be set to 50 %, 30 %, and 20 %. However, if the fitness function was shifted upwards so the fitnesses were 25, 23, and 22 instead, then this would lead to selection probabilities of 35.7 %, 32.9 %, and 31.4 %. This shows the problem with fitness-proportional selection: merely offsetting and not otherwise altering a search landscape changes the search dynamic drastically. This doesn't change much if the fitness values are divided by the mean fitness as often propagated for traditional GAs. Renormalisation procedures are generally necessary of which there are many implementation possibilities. An efficient solution is to base the selection probabilities upon the ranking in the population as demonstrated in the steady-state GA approach by Whitley et al. [507, 509].

So far only real-valued and binary genotype encoding schemes have been discussed. But binary encoding is just one special instance of the more general case of integer encoding.[40] For some real-world problem integer encoding will look like

[40] And integer encoding can in turn be generalised to allow a variety of arbitrary symbols for each gene, not only numbers. An example is genetic programming (GP).

the most straightforward encoding scheme. Think of a bridge where the numbers of columns, trusses, and steel ropes have to be optimised, or job shop scheduling problems. Integer encoding can be regarded as encompassing the intermediate cases between binary and real-valued encodings. This author believes, however, that the intermediate cases are dominated by the union of disadvantages of both extremes, rather than by the union of advantages. One gives up the intrinsic well-behavedness of a highly compressible real-valued objective function without really receiving the full information generation power a CO operator can have when acting on binary chromosomes. Remember that two consecutive one-point CO operations or one two-point CO, where the result is the alteration of one or a few bits, are equivalent to flipping these bits. But this feature relies heavily on the notion of gene pool diversity. It works only if for the concerning loci on the chromosome the population holds all the possible letters. If there are just two letters "0" and "1", then it is easy for the mutation operator even at a low rate to ensure their availability. But if five, ten or more symbols have to be present for the i^{th} locus, and if a similar degree of gene pool diversity is needed for several loci, then the curse of dimensionality requires very quickly quite unrealistically large population sizes. The more symbols there are, the more severe the competition becomes between the two goals of keeping up gene pool diversity and at the same time accumulating high-quality chromosomes.

Staying with the binary representation, GAs offer an inherently easy application to problems with mixed phenotype representation. Think again of optimising a bridge because it can be a good example of a *mixed integer optimisation problem* if integer parameters (number of pillars, number of ropes) have to be tuned at the same time with real-valued parameters (height of rope-holding pole, rope thickness, rope attachment positions). A binary genotype representation can be easily devised where integer parameters are encoded with fewer bits and real parameters with a few more.

Summarising the thoughts on simple traditional GAs

Consequently, the following points should be kept in mind when experimenting with a simple traditional GA and in particular when applying it to a continuous domain optimisation problem:

- Traditional GAs when applied to real-coded chromosomes (vectors) are driven by component-wise recombination operators. This results in a low efficiency for nonseparable problems.
- When smooth and well-behaved objective functions are funnelled through a binary genotype encoding, part of the advantageous properties of the search topology will be lost. Before application it should be considered that the advantages of using a BCGA may be offset by the loss or deterioration of guiding information carried by the objective function.
- The balance between exploration and exploitation is often difficult to control. Both, BC- and RCGAs can suffer from successful chromosomes dominating the genotype population and diminishing gene pool diversity too early along the search. This is called *premature convergence*. Depending on the properties

of the search landscape, premature convergence can also occur under moderate selection pressure.

- A GA can be trapped in *stagnation points*. This is when the progress stalls at points which are no local optimum, but which can only be escaped through improbable combinations of at least two mutations (this is the case for narrow diagonal valleys, see figure T.11). An RCGA is much more prone to stagnation than a BCGA with equal (gene-wise) mutation rate because in a BCGA also CO can change the represented value of a gene if the gene pool diversity is sufficient.

- Traditional RCGAs, in treating the search vectors component-wise, ignore most of the geometric information on the objective function that can possibly be gained; and the geometry of all moves made in the search space are not only dependent on the objective function, but also on the coordinate system of the genotype representation.

For discrete combinatorial problems, the important considerations are surely different ones. Discussing the objective functions, the notion of compressibility stays the same, but where in the continuous domain one speaks of piece-wise smoothness and valley shapes and patterns, the terms of *symmetry* and *neighbourhood* are discussed instead in discrete problems.

Binary encodings were decided not to be of interest in this work. Therefore, the rest of this chapter will only deal with real-valued encodings. Readers with further interests in EAs working on binary representations should perhaps undertake to investigate whether or when there is some meaning to the *schema theorem* and the notion of *implicit parallelism of GAs* [209] and how these things should go together with the curse of dimensionality, the eventual nonseparability of the real-world problem of one's own interest, and the no free lunch theorems. A short introduction with good illustrations to the schema theorem can be found in [507], and some assessments here [203, 219, 492].

The above description of traditional simple GAs provokes of course a question about modern and more sophisticated GAs, but the problem is that the field is a marketplace with infinitely many ideas but none of dominating popularity. The two GAs referenced and benchmarked recently together with modern EAs by Derrac et al. [114] may constitute a representative choice of good papers to start with. The next sections point to some implementation details which are of interest in modern GA versions.

Real-coded GAs

If one is aiming at solving continuous domain engineering and design problems, and if it is deemed that the shape of the real-domain objective function holds valuable guiding information, so the choice would fall on a RCGA, then it would not be advisable to stick to a traditional GA setting. Many RCGAs have been developed showing substantial improvements with respect to the shortcomings described above. In principle, there can be modifications in the three areas (a) enactment of selection

pressure, (b) mutation schemes, and (c) recombination schemes. Areas (b) and (c) are of particular interest in real-coded GAs.

a) Selection pressure. Classically, an individual's reproduction probability is determined proportional to its absolute or relative (within the population) fitness value. One simple deviation is to base reproduction probability on ranking instead of fitness. A next step could be a combination of elitism (the best few parents are copied unchanged into the next generation, that means they are being conserved) with relaxed selection pressure and higher mutation rates (more exploration, less exploitation) for the rest of the offspring [280]. But there are also more subtle measures [130] to support gene pool diversity. E.g., after ensuring that a couple of the best solutions appear among the offspring, an additional criterion for offspring generation might be to enlarge the euclidean (or Hamming[41]) distances among the set of potential parent chromosomes.

b) Mutation schemes. For real-coded chromosomes the mutation of a gene, which is a floating point number, always consists of the addition of a random number of a certain probability distribution. The distribution's shape and width offer wide space for modification and the implementation of step size adaptation schemes. An interesting example is the BGA mutation operator [314], invented to avoid the danger created by mutation step size adaptation of narrowing down too quickly and neglecting exploration. It is explained in the following. Be the interval $[-A, A]$ a domain containing a uniform distribution of mutation steps, and let it cover in each dimension a substantial fraction of the search space. For each single mutating individual (but not for each mutating gene) this interval is to be shrunken by multiplying it with a factor 2^{-k} where all k from 0 to 15 are equally probable. The wide base interval $[-A, A]$ is assumed to be left fixed during many generations of the search. This leads to a distribution favouring close neighbours down to very fine resolution without ever giving up the possibility of far mutations allowing to cross the whole search space. It is a compromise giving up on theoretically optimal population progress rates, but it yields the two benefits of no necessary step size tuning and increased gene pool diversity which made this operator somewhat popular. A larger deviation from traditional GAs is made when implementing mutation not component-wise but as the addition of vectors where the mutation vector-generating distribution is related to the search space and the state of the EA and the population.

c) Recombination schemes. Diverse literature [114, 119] shows that the BLX-α operator can make a modern RCGA competitive. This operator does not only sample a parent subspace but can also extrapolate vector differences found in the population cloud of genotypes and can hence enforce deviations from a spherical shape, i.e. contribute to stretch a deformed cloud even farther. This can come close to probing for first order parameter correlations. We will see that two other competitive EAs, DE and CMA-ES, are very similar in this respect. Recombination operators will be covered in more detail below.

[41] For any two chains of symbols of equal length, the Hamming distance is the number of differing entry pairs.

T.4.4 The recombination operator

Let us recall again the function of recombination and crossover in the evolution of genetic systems. The simplest form of MNEA is repeated cloning with intermittent mutations as can be observed with bacteria[42]. Let us use that paradigm for a simple thought experiment. If two certain point mutations are necessary to adapt to a new environment, they have to occur consecutively within one line of descendants. An initial bacterium branches up over many generations into millions of descendants. In one of these one of the necessary mutations occurs, and this bacterium again has to branch up into many descendants to allow for the second necessary mutation to happen with decent probability soon enough within its own lines of descendants. The descendants of that individual can now conquer the new environment more successfully. Now let us introduce the realistic assumption that due to the limited resources supplied by the environment the bacteria population does not just grow exponentially but stays constant. Under this circumstance it makes a big difference whether one of the two mutations alone already brings up the fitness half the way, or whether only the conjunction of both mutations increases the fitness and one single mutation is neutral. In the first case, once one of the two mutations has occurred, there will be a force increasing its frequency in the population until after many generation cycles it will prevail over most of the population. But in the second case, the momentary number of occurrences of the first mutation showing up will do a random walk starting from 1, growing to higher orders of magnitude only with very low probability. The mutation might also disappear again completely after a few generations. One can see that a series of mutations, neutral or costly when alone, but bringing substantial benefit only in combination, will be completed only with a decent amount of luck or over wastefully many generations. With millions of bacteria in a drop of water and a generation taking 30 minutes this works well enough. But it should be brought to consciousness that more complex organisms (anything consisting of more than a couple of cells and with differentiated cell and tissue types, with slower generational cycles and occurring in lower numbers) only came into being based on a drastically improved algorithm, only after the introduction of a procedure which does not only enable the mixing of genetic information from two individuals, but enforces it systematically. If each offspring individual is created by a fusion of two germ cells randomly combining the halves of the genomes from each one of two parent individuals, then this allows to recombine differing variants of different genes existing in the gene pool of the population, which greatly enhances the rate at which synergy effects between gene variants are checked out. A branching bacterial family tree is less efficient than an interconnected web of of lines of descendants. Beneficial combinations of the right mutations do not have to occur all in one line of descendants anymore. And one has to keep in mind that for complex individuals with highly locally optimised genomes the ratio between beneficial and disadvantageous or destructive random mutations will be very small. Thus it can be deemed more efficient for complex organisms to keep the mutation rate lower and recombine genome parts than to work purely with a high mutation rate. This holds particularly when taking into account that the foremost goal imposed by the

[42] and ignoring HGT via conjugation

evolutionary game on this planet is staying alive as a species, and trying out ever new genetic code is somehow secondary to that.

But nature's own EA was even further optimised in response to the problem of slow evolution exemplified by the bacterial thought experiment above. By introducing cells with double sets of chromosomes (diploidic cells, one set of chromosomes from each parent), additional memory was created along with an expanded playing field for mutation experiments. If each building block (protein) is codified twice, one of the two gene copies is allowed to be an odd one carried along. This mechanism allows to keep neutral or mildly disadvantageous gene variants in the gene pool with low frequency, the forces expelling this piece of information from storage are weak, the gene variant can keep existing or mutating over many generations until after a change of environment or the gene itself it might suddenly become beneficial. Thus, in biological evolution sexual reproduction and diploidic genomes serve the purposes of making gene recombination possible, transforming the branching tree-like structure of hereditary lines into a web, and creating additional memory for collected information, where the redundancy also allows for freer experimental modification. Hence, sometime along the development of eukaryotes between 1 and 2.5 billion years ago, the evolutionary algorithm used by nature optimised itself with the invention of recombination and diploidicy[43] and the different EA versions have existed in parallel ever since. Also, it should not be overlooked, that within parts of the domain of bacteria a different way of recombination has been invented independently, namely bacterial conjugation[44] allowing a horizontal spread of genetic information within the same generation.

In the practical world of evolutionary computation, diploidicy plays no important role (unless for simulating biological evolution), but recombination does. To explain the first point, firstly, there are easier ways to come up with information storage solutions and measures to ensure diversity of the gene pool when designing a program code. Furthermore, real-world optimisation problems are dealing in most cases with constant (for obvious reasons) fitness functions and far lower amounts of generation cycles and levels of problem complexity as the ones having justified diploidicy in nature. The need to steadily enlarge a template-based library of genes is an aspect lacking completely when EAs are just used for parameter-tuning.

The recombination operator, on the other hand, is an essential ingredient in any but the most basic EAs. Here are some popular sample implementations of the recombination or crossing-over (CO) operator (obviously, not all of them work for both, integer- and real-coded chromosomes):

- **Arithmetic mean:** The simplest way of forming an offspring chromosome by merging information from several parent individuals is to form the arithmetic mean of the chromosome vectors. (Used in a $(\mu/\mu^+,\lambda)$-ES as explained in section T.4.1.)
- **Arithmetic CO (line recombination) in RCGA tradition:** If there are two parental chromosomes $\vec{x}$ and $\vec{y}$, there will be two offspring vectors $\vec{\xi}$ and $\vec{\zeta}$ created following $\vec{\xi} = r\vec{x} + (1-r)\vec{y}$ and $\vec{\zeta} = (1-r)\vec{x} + r\vec{y}$, where $r \in]0,1[$ is a

[43] and the accompanying invention of death, see appendix U.1

[44] See appendix U.2.

random number. $\vec{\xi}$ and $\vec{\zeta}$ will lie on the line connecting $\vec{x}$ and $\vec{y}$. One speaks of **extended line recombination** if $r \in]-\alpha, 1+\alpha[$ with $\alpha > 0$.

- ***n*-point CO:** It normally involves two chromosomes. n cutting points divide them into different segments. Every second segment is swapped between the two chromosomes (see figure T.10).

- **Uniform CO (or discrete CO, DX):** For each entry (that might be a binary bit, symbol, or gene) of the chromosome it is determined independently by random from which parent it is taken. The created vectors come to lie in one of the corners of the cuboid with edges parallel to the coordinate axes that can be spanned up between $\vec{x}$ and $\vec{y}$.

- **Binomial CO:** Generalisation of uniform CO, where the probabilities of chromosome entries to come from the different parents do not have to be equal; the name has been coined in the context of differential evolution.

- **BLX (blend CO or intermediate recombination):** This works like the arithmetic CO in RCGA tradition, just that the blending parameter $r \in [0,1]$ is different for each vector component. $\vec{\xi} = \sum_i^{n_D} [r_i x_i + (1-r_i) y_i] \hat{e}_i$. $r_i \in [0,1] \forall i$. The created vectors lie inside the cuboid with edges parallel to the coordinate axes that can be spanned up between $\vec{x}$ and $\vec{y}$.

- **BLX-α (extended intermediate recombination) and BLX-αβ:** The same as above, except $r_i \in [-\alpha, 1+\alpha]$ for BLX-α and $r_i \in [-\alpha, 1+\beta]$ for BLX-αβ. Usually $0 < \beta < \alpha$. The idea behind BLX-αβ is to extend farther beyond the better parent than beyond the worse one. This means exploiting the gradient information that can be gained from the fitness difference between the parents. For all the BLX operators, one commonly takes a uniform distribution for the r_i within their intervals.

- **WHX, Wright's heuristic CO:** Like BLX, except $r_i \in [-1, 0]$. We assume $\vec{x}$ is the better parent. Seen from the position of $\vec{y}$, the offspring is created beyond $\vec{x}$ (but again uniformly distributed inside a cuboid with edges parallel to the axes).

- **Fuzzy CO:** Similar, but the component-wise random blending parameters r_i follow the probability distribution function depicted in figure T.12.

It is clear that all but the first two CO operators in the above list are coordinate system-dependent, they make the EA act differently in a rotated coordinate system. But generally it is reasonable to assume that such a representation- or coordinate system-dependence hampers a neutral, unbiased exploration of the search space. An easy solution to implement a coordinate system-independent CO operator would be to create a local cylindrical coordinate system with the origin at $(\vec{x}+\vec{y})/2$ and rotated so both parents lie on the vertical axis and create the offspring vectors in that system. Similarly, the offspring can be created via a multivariate normal distribution (see T.4.6) rotated the same way and with the centre located somewhere on the line connecting the parents, either in the middle (analoguely to BLX-α) or more towards

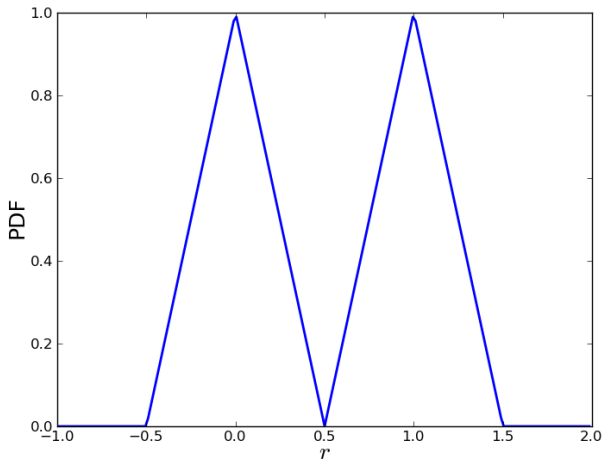


Figure T.12 Probability density function (PDF) of the random variable r used for the fuzzy recombination operator.

the better parent (similar to BLX-αβ). So the list of CO operators continues with a few selected examples from literature where coordinate system independence is ensured:

- **Unimodal Normal Distribution CO (UNDX):** This CO operator [235, 334, 335] creates two offspring from three parents $\vec{x}$, $\vec{y}$, and $\vec{z}$ by adding a mutation step $\vec{\delta}$ to the mean of the two first parents: $\vec{\xi} = \frac{1}{2}(\vec{x} + \vec{y}) + \vec{\delta}$. The mutation step is a multivariate normal (MVN) distribution[45] in the shape of an ellipsoid aligned with the line connecting $\vec{x}$ and $\vec{y}$. Along that line the σ of the distribution is proportional to the parent distance $\sigma = \alpha d_1$ with $d_1 = ||\vec{x} - \vec{y}||$, and in all other directions the distribution has a $\sigma = \beta d_2/\sqrt{n}$ where d_2 is the distance of the third parent from the line connecting $\vec{x}$ with $\vec{y}$. The second offspring is created by mirroring $\vec{\xi}$ across the centre of mass of $\vec{x}$ and $\vec{y}$. The motivating thought behind this procedure is to conserve the shape of the population cloud.

- **Simplex CO (SPX):** This CO operator [481] is the generalisation of the extended line CO for any number $m \leq n + 1$ of parents forming a simplex. The line segment between two points is a simplex covering a certain volume in the subspace defined by the two points, the infinite line going through the points. Analogously, three nonaligned points in any n-dimensional space define a 2D subspace, and the triangle in between them is again a simplex. The SPX creates offspring by taking the simplex, enlarging it by moving each point outward by a factor $1 + \alpha$, and finally sampling uniformly from the simplex' volume.

- **Simplex CO, alternative:** Xiao & Tan propose [523] to do one step of the downhill-simplex algorithm [323] with a bunch of selected parents forming a

[45] see T.4.6

simplex: find the worst one $\vec{x}_w$, construct the centre of mass of the rest $\vec{x}_c$, then construct offspring on the line between $\vec{x}_w$ and $\vec{x}_w + 2(\vec{x}_c - \vec{x}_w)$.

- **Triangular CO (TC):** This scheme proposed by Elfeky et al. [124] produces three offspring from three parents $\vec{x}_1$, $\vec{x}_2$, and $\vec{x}_3$. Three random numbers r_1, r_2, and r_3 need to be sampled with $r_i \in [0, 1]$ and $\sum r_i = 1$. The first offspring is $\vec{\xi}_1 = r_1\vec{x}_1 + r_2\vec{x}_3 + r_3\vec{x}_2$ and the two other ones are generated by cycling the indices of the parents while keeping those of the r_i fixed.

In engineering and design applications of EAs parameter ranges are often bounded and the bandwidths can differ a lot from parameter to parameter. One might have the situation $x_1 \in [1\,\mathrm{m}, 2\,\mathrm{m}]$, $x_2 \in [-0.05\,\mathrm{m}, 0.25\,\mathrm{m}]$, and $x_3 \in [180°, 270°]$. So one has to check whether CO operators based on geometric operations still act as intended even in the case of search spaces with aspect ratios far from 1. Depending on the EA and its ingredients it might be advantageous to rescale the coordinate system along some or all axes, e. g. such that $x_i \in [0, 1]\ \forall\ i$.

Summary on GAs and comparison with ESs

The underlying thought behind evolution strategies is probing the average gradient with the population cloud and doing local search with it. The idea of genetic algorithms is to explore vast areas of the search space at the same time, to have the individuals explore different distant corners, and to use the power of recombination to bring the results (parameter subset combinations) from the search agents together. ESs don't care about gene pool diversity; in each new generation the population is anchored at one single point in space. Mainly mutation drives the search by exploration starting from each new anchor point. Thus the distribution of mutation steps is of crucial concern in ESs. If a recombination operator is used, then it is only for the determination of the new anchor point by computing the centre of mass of the individuals selected. The theorising on GAs, on the other hand, is completely based on the notions of gene pool diversity and the effects of various recombination operators. Before GA application, however, one should halt and think about whether the underlying assumptions behind the special choice of heuristics are justified at all, and whether the ingredients and the setup of the GA suit the topology searched. If the problem is separable, then the traditional component-wise CO operators can be expected to do their job, no matter what representation is used. But in the separable case the choice of an EA is questionable because there are cheaper ways of solving the subproblems consecutively. The phenomenon of stagnation points shows that a real-coded representation in connection with component-wise CO operators is the wrong setup for nonseparable continuous-domain problems. In RCGAs such problems can be overcome by choosing the right CO operators and mutation schemes. The other way to go is BCGAs. Here, one gives up, to a certain degree, the ability to exploit the eventual smoothness properties of the fitness function. In return, the discretisation can be used to control the size of the search space by tuning the resolution, and the CO operators get explorative power.

T.4.5 Differential evolution (DE)

Based on the idea to use difference vectors found in the population rather than random distributions to create offspring individuals, the invention of Differential Evolution in 1994 by Price and Storn [440] yielded a simple but efficient basic evolutionary algorithm concept which has been enjoying wide popularity [101] and motivated researchers to come up with algorithm variants which have consistently proven their competitiveness among state-of-the-art evolutionary optimisers during recent years [238, 363, 364][46].

The basic idea of DE is to use difference vectors as mutation steps and also GA-style CO operators. The idea of adding difference vectors has already been described by Bremermann et al. [55] who used the operation as part of a search routine with a rather local focus after also having been examining a GA. But Price and Storn put the two things together. The DE scheme of joining the difference vector heuristic with GA elements yields an algorithm which is very competitive on global multimodal search tasks. The algorithm kernel is shown in figure T.13.

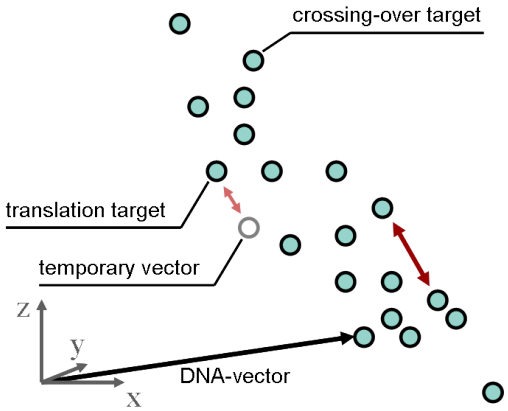


Figure T.13 The principle of differential evolution (DE).
The point cloud symbolises a population of genotypes, where the chromosomes are seen as vectors. DE does not apply mutation steps into random directions, but uses difference vectors found in the population to add them to existing chromosomes. On the lower right such a randomly chosen difference vector is shown as a long bold arrow. This vector, shrunken by a factor a, is added to another randomly selected individual, called translational target here, to create a temporary new vector, the new mutant (shown by the shaded arrow and the shaded point). Lastly, a crossing-over target has to be chosen to get mixed with the mutant. The result of that recombination is evaluated, and it will replace the CO-target in the next generation under the condition that it is better. For one generational step, this scheme has to be repeated N times, so each member of the parent population is put into question once by this scheme.

The kernel of the DE procedure can be described[47] as consisting of a mutation step, a CO step, and the acceptance rule. For the mutation step a difference vector,

[46]The two biggest conferences on EC are the IEEE Congress on Evolutionary Computation (CEC) and the Genetic and Evolutionary Computation Conference (GECCO). In recent years both conferences held competitions on real-parameter optimisation with EAs. CEC competition results of the years since 2005 can be directly accessed here: [443]. Results of the Black Box Optimization Benchmarking (BBOB) at GECCO since 2009 are to be found here: [192]. The results of these competitions are a convenient and relatively objective indicator of the most competitive state-of-the-art EAs.

[47]This description does not always abide to the nomenclature of the original publication [440] because it has been tried to come up with a more intuitively understandable set of terms.

found between two randomly chosen elements of the population, is stretched by a factor F, then added to a third element, the translation target, creating a temporary chromosome, the mutant. In the second step a recombination of this temporary chromosome with a fourth element of the population yields a new solution candidate to be evaluated. The acceptance rule is very simple: should the objective function value of the new candidate turn out to be better than the score of that fourth selected element, the crossing-over partner, then the latter is to be replaced by the new candidate in the next generation. That fourth element, which first serves as CO target and is then compared to the new solution candidate and whose replacement is at stake, is called the target of that iteration. In order to complete one generational step, during N cycles each chromosome is taken as target once, each time the new candidate vector is constructed based on a new random choice of three more chromosomes (all four nonidentical) of the parent generation and the execution of the mutation and the CO step. Each winner of such a duel will be copied into the offspring population. The algorithm pseudocode 9 shows the procedure with the following nomenclature: the target chromosome is called $\vec{x}_j$, the three other randomly chosen members $\vec{x}_{r1}, \vec{x}_{r2}, \vec{x}_{r3}$, the temporary vector $\vec{v}_j$, and the CO result, the candidate that finally gets evaluated is $\vec{u}_j$.

Algorithm 9: The differential evolution algorithm (version "DE/rand/1")

```
g ← 0
P ← GenerateRandomPopulation(N)
P ← EvaluateMembersOf(P)
while g <= G and not StopCriterion() do
    for j ← 1 to N do
        x_r1, x_r2, x_r3 ← RandomSelectNonIdentical(P,3)
        F ← rand(F_min, F_max)
        v_j ← x_r3 + F(x_r2 − x_r1)
        u_j ← CrossingOver(x_j, v_j)
        Evaluate(u_j)
        if u_j better than x_j then
            x_j ← u_j
    g ← g + 1
```

Classical DE uses two alternatives for recombination. The more common operation is "binomial" CO, which is nothing else than uniform CO steered by a probability $CR \in [0, 1]$. That means the components u_i of the candidate $\vec{u}_j$ are independently and randomly taken from the temporary chromosome $\vec{v}_j$ with probability CR and from the target vector $\vec{x}_j$ with probability $1 - CR$. The alternative is named "exponential" CO in DE literature and is a form of two-point-CO (as shown in Fig. T.10). Two cutting points are needed to compose the new candidate's chromosome. The starting index from where on vector entries are to be copied from $\vec{v}_j$ is determined randomly. One gene is copied in any case, and before each next vector component is copied, a new random variable has to be evaluated: if it is greater than

CR only once, then copying from $\vec{v}_j$ stops and the rest of the vector components are taken from the target individual $\vec{x}_j$.

Here we see that in the two cases the same setting of the strategy parameter CR may yield very different rates of incorporating new DNA, particularly for higher search space dimensionalities [533]. In the case of binomial CO CR really reflects the CO rate, so $CR = 0.9$ means on average 90 % of the candidate's chromosome is taken from the mutant $\vec{v}_j$. But in the case of exponential CO, e. g. $CR = 0.5$ does not at all mean that it is probable that half the chromosome comes from each CO member. After each component taken from $\vec{v}_j$ the copying breaks off with a 50 % chance, so for long vectors just a very little percentage of information will be taken from the mutant $\vec{v}_j$.

A shorthand was coined for describing algorithm variants. The just discussed two alternatives are abbreviated with "DE/rand/1/bin" and "DE/rand/1/exp" where the last marker tells the CO strategy, "rand" means the translational target will be selected by random, and "1" means just one mutation step is added to the translational target. This notation becomes useful when introducing other popular variants of the algorithm:

1.	DE/rand/1:	$\vec{v}_j = \vec{x}_{r3} + F(\vec{x}_{r2} - \vec{x}_{r1})$
2.	DE/best/1:	$\vec{v}_j = \vec{x}_{best} + F(\vec{x}_{r2} - \vec{x}_{r1})$
3.	DE/current-to-best/1:	$\vec{v}_j = \vec{x}_j + F(\vec{x}_{best} - \vec{x}_j) + F(\vec{x}_{r2} - \vec{x}_{r1})$
4.	DE/rand/2:	$\vec{v}_j = \vec{x}_{r5} + F(\vec{x}_{r4} - \vec{x}_{r3}) + F(\vec{x}_{r2} - \vec{x}_{r1})$
5.	DE/best/2:	$\vec{v}_j = \vec{x}_{best} + F(\vec{x}_{r4} - \vec{x}_{r3}) + F(\vec{x}_{r2} - \vec{x}_{r1})$

In variant 2 only the best among the parent individuals is used everytime as translational target building on the intention of further pushing the good end of the population cloud. Variant 3 first shifts the translational target towards or past the best point on their connecting line before applying a second mutation step. Variants 4 and 5 only differ from 1 and 2 by the application of a second differential mutation step to each translational target. The scalar F is often randomised within a given interval ([362] p. 79).

A problem of DE is that the various strategy and parameter choices yield algorithms with substantially different performance, and this also depends on the objective function topology. Hence, much effort has been put into coming up with efficient adaptation schemes. One example is saDE (self adaptive DE) which scored rather robustly in the CEC-2005 competition [193, 374] and also in a comparison of "well-known evolutionary and swarm intelligence algorithms" by Derrac et al. of 2011 [114]. These good results were achieved even though there was just one common CO rate variable CR implemented acting on both, exponential and binomial CO operators which seems to be a suboptimal implementation. Some newer approaches are presented here: [58, 59, 374, 536].

Summary: The basic concept is constructing mutation steps based on difference vectors found in the genotype population and adding a GA-style CO operator to the mix. The mutation strategy aims for a similar effect as seen in CMA-ES (see next section) of amplifying existing deviations of the population cloud from a spher-

ical shape. The motivation is that the replacement scheme, following a greedy[48] algorithm, stretches the population cloud into the direction of the steepest average gradient and forces its progress in that direction. The influence of the GA-style CO operators used in the algorithm must certainly have a diluting effect on the population dynamic created by that mutation and selection strategy. But this diluting element seems just about the right thing to do to tune the balance between exploitation and exploration well. It has to be kept in mind that the commonly employed CO operators introduce a certain degree of coordinate system dependency.

T.4.6 CMA-ES

The Covariance Matrix Adaption Evolution Strategy [196] ranks among the most efficient, robust, and problem-independent evolutionary search engines available today [101, 114, 193]. It is a workhorse needing no tuning of strategy parameters in most application cases. Although its basic concept is that of a local search engine because the moving population cloud is used as a gradient probe, the algorithm shows astounding robustness against local optima when applied to highly multimodal and high-dimensional test functions [193]. The core of this ES is a special way of tuning the mutation step distribution. It consists of performing a *Principal Component Analysis (PCA)* also named *Karhunen-Loève (KL)* decomposition on the set of recently successful mutation step vectors and using the gained information for rotating a *multivariate random distribution* for the creation of new mutation steps into the right direction.

Multivariate Normal Distributions

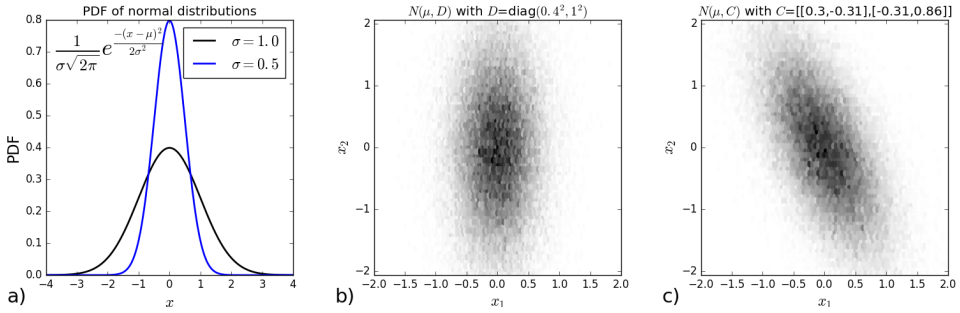


Figure T.14 The multivariate normal random distribution.
(a) Probability density function (PDF) of two normally distributed random variables with different standard deviations σ. (b) Random distribution of points in two dimensions (2D histogram). First and second coordinates of the points follow different normal distributions with means m_1, m_2 and standard deviations σ_1, σ_2. The multivariate normal random distribution is defined so the point distribution can be expressed as $\mathcal{N}(\vec{\mu}, \Sigma)$, where $\vec{\mu} = (\mu_1, \mu_2)$ and $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$. (c) The general case allows for scenario (b) to be rotated in any direction, in which case Σ is a positive definite matrix and can have nonzero off-diagonal elements. Σ is then the covariance matrix of the resulting point cloud.

In probability theory and statistics, the multivariate normal (MVN) distribution is a generalisation of the one-dimensional (univariate) normal distribution to

[48] The *greedy algorithm*: create variations, never reject an improvement, never accept something worse than the current (set of) solution(s).

higher dimensions able to exhibit multiple variances in different directions. Let $\vec{x}$ be an n-dimensional random vector, constructed by sampling component after component from the same one-dimensional normal distribution and its probability density function (PDF):

$$\mathcal{N}(\mu, \sigma^2) \quad \text{with PDF:} \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}. \tag{T.5}$$

The resulting point cloud will be centred around $\vec{\mu} = (\mu, \mu, ..., \mu)$ and density isocontours of the distribution will be spheres. This distribution can be expressed as

$$\mathcal{N}(\vec{\mu}, \sigma^2 I) = \vec{\mu} + \sigma\mathcal{N}(\vec{0}, I) \tag{T.6}$$

where I is the identity matrix. Two example PDFs of normal distributions are shown in figure T.14 (a). After modifying the above procedure in a way that each component x_i of a new random vector $\vec{x}$ will be generated according to an own normal distribution with own values for mean and variance, μ_i and σ_i, the resulting point distribution will be centred at $\vec{\mu} = (\mu_1, ..., \mu_n)$ and stretched or compressed along the coordinate axes, as shown in figure T.14 (b). The density isosurfaces will be ellipsoids aligned with the coordinate axes. The notation allows in this case

$$\mathcal{N}(\vec{\mu}, D^2) = \vec{\mu} + D\mathcal{N}(\vec{0}, I) \tag{T.7}$$

where $D = \text{diag}(\sigma_1, ..., \sigma_{n_D})$ is a diagonal matrix. In the general case where the density isosurface ellipsoids need not be aligned with the coordinate system, the second argument of $\mathcal{N}$ must still be a positive definite matrix and the covariance matrix of the resulting point cloud will tend towards that matrix if the number of points tends towards infinity.

$$\mathcal{N}(\vec{\mu}, C) = \vec{\mu} + C^{\frac{1}{2}}\mathcal{N}(\vec{0}, I) \tag{T.8}$$

Principal component analysis (PCA)

Assuming we have any irregularly shaped point cloud, the *Karhunen-Loève* decomposition or decomposition into *principal components* is a procedure to devise a new orthogonal coordinate system such that the first coordinate axis coincides with the direction of the widest variance of the point distribution, and that from the 2nd to the n^{th} coordinate axis the variances get ever smaller and whereby every new coordinate axis has to be orthogonal to all other ones. It turns out that the eigenvectors of the covariance matrix of the distribution, in the order of decreasing eigenvalue magnitude, provide a satisfying set of orthogonal basis vectors. Figure T.15 shows some examples for distributions and their principal components. In the case of a multivariate normal distribution, the PCA finds basis vectors for a coordinate system in which the density isosurface ellipsoid is aligned. This can help reducing the problem formulation to the small subspace of interest. In the case of a straight, thin, needle-like distribution of points, the 1st PC gives the direction of the line. That vector allows for a one-dimensional description of the problem, giving the simple linear correlations between the coordinates, which restrict possible points to positions

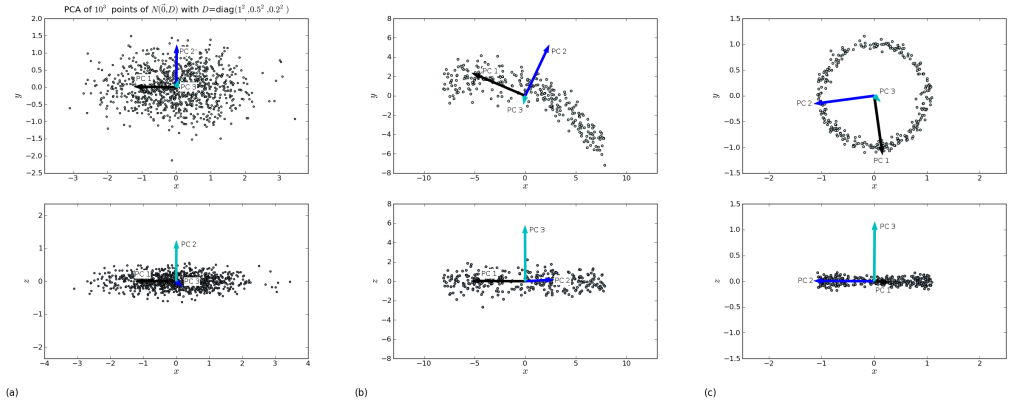


Figure T.15 PCA of diverse point clouds.
(a) Applied to a multivariate normal distribution, the PCA finds the density iso-ellipsoid's axes. Where points are distributed along straight lines with slight noise, PCA can uncover the line's direction and give a good approximation for the linear correlation which allows to reduce the problem size to one dimension. (b) This data set is uniformly distributed along x, and normally along y and z with same σ. Finally, the points have been shifted along y depending on the x-position which gives the cloud a banana-like shape. The 1st and 2nd principal components coincide nicely with the xy-plane, in which the banana lies, the 1st PC giving the direction of largest variance and the 2nd PC being oriented orthogonally to the 1st one. (c) This circular point cloud illustrates an example, where the PCA is not the appropriate means to uncover a simple functional correlation able to reduce the size of the problem to one dimension.

along a line with some noisy offset. The other PCs of the needle-shaped cloud will have eigenvalues of marginal magnitude compared to the eigenvalue of the 1st PC. In general, the PCA can be used to determine which subspace (line, plane, or hyperplane) is relevant for the problem dynamic using the magnitudes of the eigenvalues as truncation criterion. PCA can be made part of a process for formulating reduced-order models (ROM) based on statistical data sets generated from complex systems. Figure T.15 (c), however, illustrates the limits: for this distribution of points along a circular line, the PC eigenvalues will clearly state that the problem mainly plays in the xy-plane, but PCA will be unable to uncover the one-dimensional nature of the problem because the covariance matrix as the central key of the approach only probes for linear correlations.

CMA-ES concept

The CMA evolution strategy is basically a $(\mu/\mu, \lambda)$-ES, often with $\lambda = 2\mu$. That means that in each generation the best 50% of individuals are selected, the mean of all those chromosomes is calculated, and all offspring individuals are created by applying a mutation step to that mean vector. Now it is all about how the mutation operator looks like in detail, i. e. what shape is given to the probability distribution of the mutation steps. The idea is to use a multivariate normal distribution tweaked the right way, so it suits the situation currently encountered by the algorithm, that means to use the information available from the data of the most recent generational steps to get an approximation of the average gradient in the area covered by the population and turn the distribution of mutation steps into the ideal direction shown in fig. T.9 on the right. This is done by performing a PCA on the subset of the best

μ mutation steps of the past generation and use the covariance matrix gained to tune the MVN distribution of mutation steps accordingly for the generation of the next offspring. The two main internal state variables of the algorithm are that covariance matrix C and a global mutation step size parameter σ. There is some inertia introduced by exponentially smoothing the development of those two state variables, meaning that there are additional recursive terms in the updating formulae for C and σ. Hence, the history of the evolution path is also taken into account. The entire concept was published by N. Hansen and coworkers in a series of papers [18, 196, 197]. Sample program codes are available online [191]. The algorithm's pseudocode is given below.

Algorithm 10: The CMA-ES

```
g ← 0                                                  // generation counter
P_F0 ← GenerateEmptyPopulation(μ)                      // parent population
P_F1 ← GenerateEmptyPopulation(λ)                      // offspring population
x_a ← x_0                                  // initial anchor point x-start
σ ← σ_0          // initial setting of mutation step size parameter
C ← I                    // initial setting of the covariance matrix
while g <= G and not StopCriterion() do
    P_F1 ← MutateMVN(x_a, C, σ)
    EvaluateMembersOf(P_F1)
    P_F0 ← Select(Sorted(P_F1), μ)
    x_a ← WeightedMean(P_F0)
    C, σ ← Update(P_F0, C, x_a, x_a,old)               // recursion -> history
     influence
    g ← g + 1
```

It is interesting to view CMA-ES as the result of a long development process in the ES community: in the historic evolution of ES the mutation distribution is initially just controlled by one scalar step size parameter σ, thus the distribution is isotropic. Next, n independent strategy parameters σ_i are implemented. The mutation distribution gets the shape of an ellipsoid aligned[49] with the coordinate axes. Finally, introducing yet more control parameters for rotating the distributions allows to break the alignment [21, 406]. Therefore, the idea of seeking suitably stretched and rotated MVN mutation distributions does not represent the novelty of CMA-ES. That lies rather in the way the adaptation works. In Schwefel's algorithms of 1979 [21, 406] all the additional strategy parameters are hereditary attributes of individuals subject to mutation which are expected to be adapted by evolution to suit the encountered search landscape. But the standpoint taken by Hansen, Ostermeier, & Gawelczyk of 1995 [197] is different, and it follows the same principle as the $1/5^{\text{th}}$-rule, the principle of transcending the abstraction of MNEA by tapping into the power of the swarm. If the statistical information is available, why not use it immediately? If done the right way, the positive effect can be realised much quicker as compared to adaptation schemes based on evolutionary gene pool improvements

[49] See figure T.9 and accompanying explanations (page 516).

where the intended effects arise with a delay of many generations.

Why and how CMA-ES works

CMA-ES is the response to the motivation represented in figure T.9 (p. 516) expressing thoughts on the ideal setting of a local search engine. But how can this paradox be understood that CMA-ES is by design a local search engine, and yet shows such a competitive performance when used as a global optimiser? Two thoughts are offered in response to this question which are hoped to be instructive for understanding some basic features of population-based search engines. The two thoughts are expressed in two figures.

The first one, figure T.16, proposes to view an ES as a dancing spider subject to a shrinking spell. The spider initially is of planetary dimension, somebody has cast a nasty shrinking spell on it, so in a few minutes it will be of normal spider size, and the task of the spider is to make sure it ends up sitting on one of the highest mountain summits on earth by that time. For that purpose the spider can probe the altitude of the landscape with its legs (it is just blind) and can move sideways to where it thinks to have found the most promising mountain chains. Therefore it will look like the spider is dancing while it is trying to get as much data as possible in the limited time span. It must be wise in its probing strategy and, most importantly, in its decisions on sideways movements. Say, it starts out sitting above the North Pole with its legs being able to reach all of North America and Eurasia. If it decides too impulsively, it might move over to America very soon after one lucky random sample of the Rocky Mountains, and the Himalaya will get out of reach when the spider has begun shrinking. If it decides and moves too slowly, then it will miss the Himalaya as well, because after some shrinking with little or no sideways motion, the much denser statistics from Greenland and Europe will overrule any memory of eventual initial samples from Tibet. Later on, a reluctance to decide in time between the Pyrenees and the Alps might cause the spider to end up on the less high Massif Central. For that purpose, CMA-ES is equipped with an elaborate scheme, recursive formulae for updating σ and C aimed at balancing the influences of memory content and new information input. In historic ES literature the symmetry of the implemented step size adaptation mechanism is often stressed. It was assumed that the spider should be able to grow as easily as it can shrink, which makes sense under the assumption of local search, i. e. under the conditions expressed in fig. T.8 (p. 515). But the modern implementations of CMA-ES, DE, and PSO codes taking part in benchmarking competitions (e. g. at GECCO or CEC conferences) bear witness that in the discipline of global real-parameter optimisation a growth ability for the area covered by a population is obsolete, a growth rarely ever happens in practice. With these EAs the most intelligently guided shrinking process is the goal. After beginning the search broad and globally, the question is how fast to zoom in and towards which spot.

The second thought is expressed in the sketches and histograms of figure T.17 and it says that by using CMA-ES the spider in fact tries to get an idea of the envelope of the mountain summits. The top row diagrams in the figure show 1-D search landscapes (the bold cyan curves) which, if treated as minimisation problems,

Figure T.16 Inner workings of CMA-ES, part I: the spider.
Why is CMA-ES such a competitive global search algorithm even though by design it is a local search concept? This can be understood by imagining the search engine as a dancing spider subject to a rapid shrinking spell. The spider has the task of ending up on the highest mountain summit by the time it reaches normal dimensions and it is allowed to probe the surface with its dancing legs. The spider firstly needs the right probing strategy. But more importantly, it needs the right decision strategy for moving sideways: it shouldn't move too lazily because shifting the search focus to another continent is only possible while the spider still is of a size comparable to continents, but it also shouldn't jump too quickly to the first high mountain felt on one side without backing the decision up with a bit of statistics from the other side. (Artwork: Tudor Pirvu)

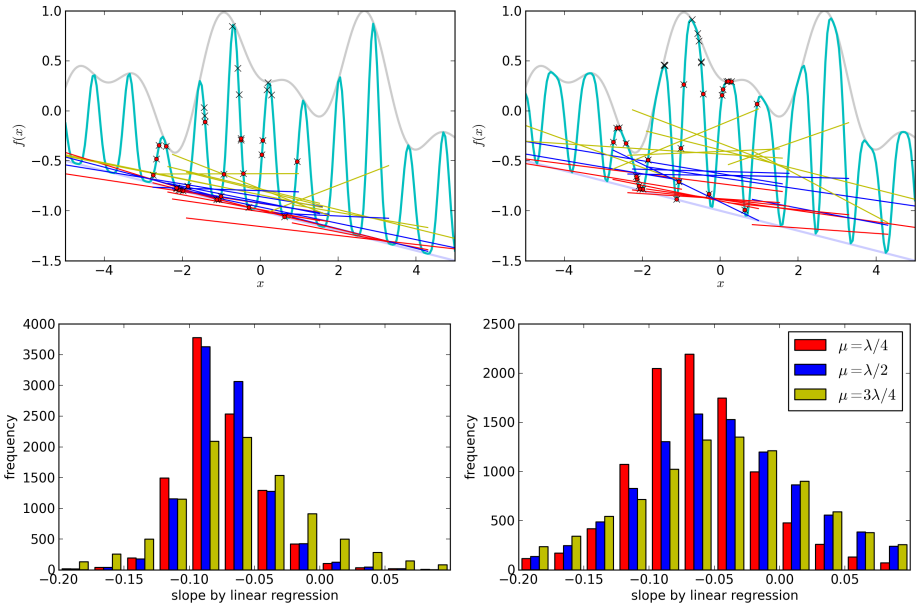


Figure T.17 Inner workings of CMA-ES, part II: the envelope.
These illustrations and statistics (described in the text) are intended to explain under which conditions a simple (μ, λ) selection strategy applied to uniformly distributed samples can lead to a data set able to capture the slope of the envelope of the valley bottoms. Note that on the left there are narrow mountains separating wide valleys and on the right it is the other way round.

contain very useful information guiding towards the global minimum. The guiding information is the surface connecting the valley bottoms, which is here a simple linear function. The wavy structure is based on the function $f : \mathbb{R} \to [0, 1]$ with

$$f(x) = \begin{cases} \left(\frac{1}{2}\left(\sin(\omega x) + 1\right)\right)^{(p+1)} & \text{if } p \geq 0 \\ 1 - \left(\frac{1}{2}\left(\sin(\omega x + \pi) + 1\right)\right)^{(-p+1)} & \text{if } p < 0 \end{cases}$$

where the parameter p controls the relation between mountain widths and valley widths. This regular wave with the frequency ω is linearily projected from the interval $[0, 1]$ to the space between the two envelope functions drawn lightly in grey, the linear function at the bottom, and an irregular function at the top. The noteworthy difference between the two cases is the setting of p. On the left $p = 1$ leads to wide valleys separated by thin peaks, and on the right, with $p = -1$, massive wide mountains are separated by narrow canyons.

In the CMA-ES the strategy parameters are updated in each generation by exclusively using information from the best μ members of the population of size λ. The black crosses in the diagrams are a set of 32 random samples distributed uniformly across a certain interval along the x-axis. The 16 best crosses are also marked by red dots, they represent a selected data set. Now one can fit a linear function $f(x) = ax + b$ to the selected data set and see how well it would represent the gradient of the surface connecting the valley bottoms. This procedure has been repeated 20 times with 20 different sets of sample points spread along 20 randomly determined intervals, all covering at least three wavelengths. The lines scattered across the plots show the results of the linear fits. There are in total 60 lines shown in the plots because each data set has been fitted three times with different (μ, λ) selection schemes. The red lines were gained by selecting the best quarter of points, the blue lines are based on selecting the better half, and the yellow lines show slopes deduced from the best three quarters. The random probing locations are the same in the two plots. When determining the selection ratio $\frac{\mu}{\lambda}$ one has to deal with the trade-off that a small ratio entails the problems of a smaller statistics and that a larger ratio leads to more influence from the higher parts of the landscape, in which one has no interest, and which often may be assumed to be more misleading than helpful. What these plots qualitatively show is that in all cases with the exception of the yellow lines on the right, the search direction (left or right?) is given almost always correctly and the search engine would be able to deduce the right way for the spider to go next. So in the case of the wide valley the choice of the selection ratio matters a lot less than in the case of narrow valleys since most points except some unlucky few lie near the valley bottoms anyway. The comparison of the red and blue lines in the wide valley case shows that there are some more outliers among the red lines which can be attributed to the smaller statistics. The fact that most lines underestimate the lower envelope's slope and almost none of them exaggerates it, is due to the simple fact that the selection cutoff is always a horizontal line, so each dataset stems from a triangular area between the lower envelope and the cutoff.

The lower row of plots in figure T.17 show the histograms of slope estimates on the basis of repeating the experiment with $\lambda = 32$ ten thousand times. The true slope of the bottom envelope is -0.1, so the tendency to underestimate the slope is

visible in all datasets. But the more important observation one can make is that in the left histogram the yellow data set is the odd one, whereas on the right it is the red one standing out of the group of three. This means, firstly, that the common choice of $\lambda = 2\mu$ is backed by this experiment as a good and safe guess. The second lesson is that a generous selection is okay in the case of wide valleys. But a landscape of narrow valleys, by contrast, requires sharper selection, and particularly in this case the CMA-ES or similar (μ, λ)-EAs will be very sensitive to the setting of the selection ratio.

Summary: Hence, the working principle of CMA-ES, one of the most competitive function minimisers existing, can be understood by assuming synergy effects between the following three features: (a) the (μ, λ) selection strategy approximates the envelope of the valley grounds, (b) the shaping of the mutation step distribution is the mechanism responsible for stretching the population cloud amoeba-like into the direction of the steepest gradient, and (c) the memory-to-input balance in the updating routines for the position and the distribution shape lets the spider negotiate wisely its antagonising desires for quick moves and reliable statistics during its shrinking journey. The spider image can explain why CMA-ES shows strong performance on self-similar landscapes with a detectable average gradient, on topologies which pose comparable challenges and similar levels of structure and ruggedness while sinking down through a wide range of zoom levels. (This is the case for the Weierstrass function, number 11 in [444].) It also should be recalled that the free and amoeba-like shaping of the population enables the CMA evolution strategy to efficiently follow down narrow diagonal valleys (fig. T.11), something that can be accomplished by downhill-simplex search [323] but not by simple GAs. Understanding the character of CMA-ES can of course help not to waste time and resources with it in application cases not suitable to this EA, which is the case if the envelope of the valley bottoms does not offer guiding information or is even misleading. An example of an irresistibly misleading search landscape for CMA-ES is the Lunacek bi-Rastrigin function, numbers 17 & 18 in [265]. In these more desperate cases CMA-ES makes only sense with small population sizes as efficient local search with many random restarts.

T.4.7 Particle swarm optimisation (PSO)

The inventors of particle swarm optimisation, R. Eberhart & J. Kennedy [120, 230], were not inspired by the evolution of genetic systems, but by the movements of swarms of birds and fish. They created small 2D simulations of moving points in a periodic rectangular plane where swarm members would be influenced by the position and vector of motion of their closest neighbours. As one goal they wanted to find the conditions under which collisionless synchronous swarm movements would arise. In a different simulation they intended to recreate the situation of birds coming from far away and beginning to swarm around a bird feeder where a whole swarm can profit from the discovery made by just one or a few members. One of their simulations became a widely used and efficient EA template, outlined below as algorithm 11.

In Eberhart & Kennedy's PSO template (algorithm 11) one member of the pop-

Algorithm 11: The particle swarm optimisation algorithm

```
P ← generate_random_population(N)
P ← evaluate_members_of(P)
for p in P do
    p.L ← assign_communicating_neighbourhood(p,P)
    p.v ← random_velocity_vector()
    p.bestx ← update_own_best_ever(p)
for p in P do
    p.Lbestx ← update_neighbourhood_best_ever(p.L)
while t <= T and not stop_criterion() do
    t ← t + 1
    for p in P do
        for i in [1, 2, ..., n] do
            p.v[i] ← α p.v[i]                                   // inertia
            p.v[i] ← p.v[i] + β rand()(p.bestx[i] − p.x[i])     // memory1
            p.v[i] ← p.v[i] + β rand()(p.Lbestx[i] − p.x[i])
             // memory2
        p.x ← p.x + p.v
    P ← evaluate_members_of(P)
    for p in P do
        p.bestx ← update_own_best_ever(p.f)
        p.Lbestx ← update_neighbourhood_best_ever(p.L)
```

ulation $\mathcal{P}$ is thought to represent a particle or agent proceeding on its own trajectory through the search space and communicating about what it sees with other members of the swarm so the shared information benefits the swarm's search performance. Unlike the individuals of normal EAs, which are solely defined by their chromosome $\vec{x}$ and the fitness $f(\vec{x})$, the swarm agents with their information processing abilities need to be slightly more complex entities and associated with more data structures. Let's describe one particle `p` as in an object-oriented code framework where short names can refer to complex data structures containing relevant data as attributes. E. g. the particle's position vector $\vec{x}$ is to be stored as `p.x`, its velocity vector $\vec{v}$ is to be named `p.v`, and the fitness `p.f`. The movements of the agents are determined by four forces: (a) inertia, (b) the own history, (c) randomness, and (d) experience communicated throughout the swarm. The strengths of the forces are tuned by two important state parameters α and β, where the first modulates the inertia effect, and the second the swarm experience forces. In the case of $\alpha = 1$ and $\beta = 0$ the particles fly straight through space at constant velocity. If β is raised above zero, but the random influence is still kept switched off, then one can imagine that each particle is suddenly anchored with a rubber band at some point in space and starts to oscillate around that *attractor point* in circles or ellipses. If α is reduced below 1 it means that the particles lose speed as if by friction[50] and the orbits around their attractors get an increased tendency to shrink and become continuously smaller, so spiralling lines are drawn. If the random force (c) is turned on, then the swarm forces will be modulated separately in each dimension by independent random numbers sampled uniformly from the interval $[0, 1]$. The question is only what the anchor points or attractors of the orbits are. The anchor points are set by the memorised experience of the swarm. In the above pseudocode it is a combination of two influences, one coming from the best point ever visited so far by the agent itself (the point `p.bestx` with fitness `p.bestf`), and the second one being the best ever spot so far discovered by the particular subset of the swarm `p.L` which communicates to the agent `p` (the point `p.Lbestx` with fitness `p.Lbestf`). Since those two influences are weighted equally, it can be imagined that the orbit centre, the effective attractor, is the point lying in the middle of the two attractors [231]. If the search is in a state where the attractors are replaced often, then it is clear that the particles will not often complete whole orbit cycles. Consequently, the trajectories will be sequences of bends into various directions. And additional kinks from the random influence will make them appear even more irregular.

The effect of the implementation of the random influence deserves some special consideration. There is an important difference between two modes of implementation: in the above pseudocode the random number generator is sampled $2n$ times

[50] As can be seen in the first part of the formula for the velocity update $\vec{v} = \alpha\vec{v}$, $\alpha < 1$ means friction and it aims at velocity decay and facilitating swarm collapse. In practice there are however regimes where even with $\alpha < 1$ particle velocities and orbit radii often increase. PSO is a global search tool and no swarm simulation engine. The time steps are chosen coarsely, the particles make large jumps in each time step and thus form kinky trajectories. The goal of the algorithm is exploration and not the drawing of smooth trajectories. Tangential orbit steps determined by the extrapolation of the latest jump by the vector `p.v` and the attractor forces will generally tend to increase orbit radii. Therefore, α has to be smaller than 1 just to stabilise an orbit, as will be explained further below.

during one agent's position update, so the two memory influences are scaled in each dimension with a different amplitude. In an alternative implementation just two random numbers are sampled and the two forces are scaled in each dimension with the same amplitude. The result of the latter approach is that as long as the attractors `p.bestx` and `p.Lbestx` stay unchanged, `p` can never leave the plane defined by itself, its speed, and the effective attractor. This reduces the searched subspace per orbit. In later phases of the search when the swarm memories do not get updated that often any more this reduction increases in severeness. In the component-wise implementation version, however, the random kinks in the orbits make `p` leave the plane. It was pointed out by Wilke [515] that there is some confusion in literature caused by authors unaware of this difference.

The last definition needed to characterise the particle swarm optimiser is the set of rules on who communicates with whom in the swarm. It has been found that too much communication decreases the performance of the algorithm on hard multimodal problems [231]. That means if every agent of the swarm is in possession of the knowledge about the best spot ever visited by any member of the swarm, i. e. the global optimum found so far, it leads to too much influence of that piece of information and to a too undisturbed collapse of the swarm into its vicinity. PSO shows better performance if each `p` can inquire only a small subset of $\mathcal{P}$ about their memorised experience. This means that for discussing PSO one has to look at two different swarm topologies and their interrelation: the topology of the agent positions in the n-dimensional search space and the topology of the graph symbolising the communication channels between the swarm members. The communication graph is implemented by giving each agent `p` a list of other agents `p.L` which it can ask about their best spot memory. `p.L` represents an agent's societal locality. `p.Lbestx` is the storage place for the best spot an agent can identify after having compared the gathered memories of its social neighbourhood. Together with its own best spot memory `p.bestx` these two attractors create the swarm force acting on the agent's trajectory besides inertia and randomness.

The most common approach for the communication topology is a von Neumann topology made by ordering the N members of $\mathcal{P}$ into a toroidally wrapped rectangular grid. Here, `p.L` may contain an agent's four next neighbours or neighbours up to a certain degree. Plenty of other topologies can be found in literature and code projects. Note that by following the original PSO template of first setting up the communication graph and then distributing the particles randomly at their starting positions in the search space $\mathcal{X}$ a situation will be created where the two topologies have nothing at all to do with each other. Communication lines between close neighbours in $\mathcal{X}$ are not favoured over connections between very distant particles. This certainly yields larger average initial orbits than would be the result of an alternative approach where the neighbourhoods `p.L` were set up to reflect neighbourhoods in $\mathcal{X}$. The assumption that a high ratio of far-distance communication channels are beneficial to the global search behaviour is supported by the findings of Liang [264] who therefore proposes a periodic random reshuffling of the communication graph.

Since the invention of PSO in 1995 many algorithm variants have been explored and published. They differ in their communication topologies, and in many aspects of the central formulae for updating particle velocity and position. Often the state

parameter β is split up into a constant c_1 modulating the influence of `p.bestx` and another one c_2 for `p.Lbestx`. Measures for supporting diversity are proposed like randomly restarting single particles once in a while, or replacing inertia with random vectors (Gaussian PSO, [242]). Other authors use PSO to build more complex optimisers on it (stretching PSO, PSO parameters controlled by DE, [343]).

There was also some effort put into understanding the original algorithm template, improving and generalising it, and giving theory-based guidelines for the algorithm's parameters. Some results are referenced in [231, 357]. A generalised PSO kernel formula for updating the speed and position of the j^{th} element of the swarm can be given by

$$\vec{v}_j \leftarrow \alpha\vec{v}_j + \frac{\psi}{K_j}\sum_{k=1}^{K_j}\vec{\mathcal{U}}(0,1)\otimes\left(\vec{\xi}_k - \vec{x}_j\right) \tag{T.9}$$

$$\vec{x}_j \leftarrow \vec{x}_j + \vec{v}_j. \tag{T.10}$$

Here, the subroutine $\vec{\mathcal{U}}(a,b)$ creates an n-dimensional vector for scaling the forces in each dimension with independent random numbers distributed uniformly along the interval $[a,b]$. The $\vec{\xi}_k$ are the attractors, and their force scaling is given by a constant ψ which has to be divided by the number of attractors K_j acting on the trajectory of agent j. If the $\{\vec{\xi}_k\}$ are the best memories of all swarm members ($K_j = N\ \forall j$ with N the population size) then one speaks of a fully informed swarm (FIPS). The original PSO template outlined above reflects the case $K_j = 2\ \forall j$, hence $\beta = \frac{\psi}{2}$. Useful and common values are $\alpha = 0.7298$ and $\psi = 2.9922$, according to [231]. If β is split into c_1 and c_2 as mentioned above, then their sum should not become smaller than 2β, according to [357]. More scenarios, parameter guidelines, and references to the underlying theoretical articles can be found in [357].

The sets of variables (α,ψ), (α,β), or (α,c_1,c_2) balance inertia against the swarm forces and are very important for stabilising the swarm so it does not explode and allowing a steady convergence behaviour, so that the swarm can "glide down from larger to smaller scales" as with CMA-ES or DE. If PSO were a realistic simulation engine for simulating a conglomerate of planetary systems where particles have potential and kinetic energies, where there are attractive and centrifugal forces, where the goal is to draw smooth and realistic trajectories, then it would seem weird to chose a value for $\alpha = 0.7298$, so that 27 % of kinetic energy gets dissipated in each time step. But the purpose of PSO is not to spend the whole budget of calls to $f_{\text{obj}}(\vec{x})$ on a few short stretches of high-resolution trajectories. The purpose of PSO is global search and explorative probing of the search space. Therefore the lengths of the jumps made by each particle in one time step must start out quite large, thus the orbits will be edgy spiralling structures. Looking at the (unrealistic) special cases of a particle orbiting around an attractor by forming a square with four jumps or a pentagon with five, it is clear that the inertia vector for the next step never points in the tangential direction, but always outwards. Even a large step into the tangential direction would increase the orbit radius, a step along an outwards pointing vector even more so. If the attractor force of the same time step does not reach far enough inwards to counteract this tendency, then the orbit radius will diverge. The state variables of PSO have the job to balance the tendencies of orbit expansion

and contraction in order to yield the slow and steady convergence behaviour desired for global search engines. That PSO is not framed in terms of kinetic and potential energies has another consequence: angular momenta are not necessarily conserved. There are time step geometries allowing under certain circumstances from the random number generator the quick step-by-step degradation of angular momentum, where a round or elliptical orbit can be transformed in a few steps into an oscillatory movement along a line (or into very much narrower ellipses). It was shown by Wilke [515] that the danger of collapse to line search is elevated a lot when using single random numbers for scaling the swarm influences equally in all dimensions instead of multiple random numbers, i. e. when using $\mathcal{U}(0,1)$ instead of $\vec{\mathcal{U}}(0,1)$.

Summary on PSO

PSO as a swarm algorithm is not described in the habitual EA terms of chromosomes, mutations, recombinations and so on. The biggest difference is that the fundamental evolutionary driving force of *survival of the fittest* seems to be lacking completely, every agent always survives into the next generation. But the application of Atmar's wider definition[51] of information generated and handed down (or searched by trial and error and spread) in competition shows how to include PSO into a definition of evolutionary systems and compare it with other EAs in this framework. In PSO the selection happens not on the level of the agents but on the one of their memories. Where old-school EAs have recombination operators PSO has implemented a topology of information channels. Discussing the influence of random number implementations in PSO is clearly the same as discussing mutation schemes elsewhere. The feature of the PSO algorithm that sets it apart from other EAs is that attractors don't just spread their chromosome by senseless copying. Instead, every trajectory towards an attractor is deliberately designed to miss its target and to fly by at a finite distance.

How about the swarm's ability to follow down narrow diagonal valleys, i. e. to deal with squeezed, stretched, and rotated search spaces? The basic swarm and orbit laws are coordinate system-independent and can be stretched and squeezed diagonally, only the component-wise randomisation operator will constantly try to break out of such a subspace. But a GA-like stall with no way out is not part of the pattern.

PSO has the instrument of communication graph topology to play with, a concept not present in other EAs. Bounding the level of communication increases the robustness of global search. This can be easily explained by noticing that a fully informed particle swarm (FIPS) exhibits just one attractor at a time whereas a locally communicating PSO is permanently influenced by many of them. The many effective attractors are linear combinations of the raw attractors, the individuals' best search point memories.

PSO has led to remarkably efficient optimisers and algorithms of that family have been persistently present in the EA competitions at the CEC and GECCO conferences of the recent years. PSO algorithms are just one of several success

[51] see quote on page 496

stories in the family of swarm algorithms which hosts also *ant colony optimisation (ACO)*, *bee colony optimisation, (BCO)*, and many more.

T.4.8 Scatter search (SCS)

Scatter search goes back to an initial idea published by Glover in 1977 [171] and further contributions by him and a small group of researchers [170, 172, 249, 287] starting in 1994. The concept started out as a work on discrete combinatorial problems (like TSP or JSSP). In this algorithm class new trial solutions are formed by combining information from groups of two or more parent individuals. A local search heuristic is applied to each new trial. This classifies scatter search as a hybrid EA, and sometimes it is also called a memetic algorithm (see section T.4.9). Whether a trial solution will be part of the next generation's parents does not only depend on the fitness function but also on how much an individual's inclusion enhances the diversity of the parent set's gene pool. Whereas in many common EAs random numbers are used extensively in the parent choice and chromosome manipulation processes, the scatter search tries to be puristic about randomisation and to accomplish as much as possible by deterministic rules. Fred Glover justifies this purism in [172] with the following argument:

> A "foolish mistake" incorporated into a deterministic rule becomes highly visible by its consequences, whereas such a mistake in a randomised rule may be buried from view – obscured by the patternless fluctuations that surround it. Deterministic rules afford the opportunity to profit by mistakes and learn to do better. The character of randomised rules, that provides the chance to escape from repetitive folly, also inhibits the chance to identify more effective decisions.

Therefore, random processes are introduced only where they can substantially benefit the genome diversity and under the condition that it cannot be assumed that they threaten to slow the search by introducing dissipation into the otherwise more stringent workings of the particular search concept implementation. This can be seen in particular in the offspring generation scheme. In many other EAs there is a fixed target number of offspring solution candidates to be generated and a more or less biased random process chooses parent subsets and generates new candidates from the subsets until the target number for candidates has been reached. As long as the target number of offspring is not larger than the number of parents by orders of magnitude, this will always create substantial fluctuations of parent influence on the offspring gene pool. The SCS approach, by contrast, enforces an egalitarian influence by each member of the set of μ parents through keeping their number small enough that (in the idealistic extreme case of the algorithm) all the possible subsets (i.e. parent pairs, triples, quadruples, and higher N-tuples up to $(\mu - 1)$-tuples) can be worked off. Of course, in most application scenarios this extreme scheme drives the algorithm execution to become infeasible if the parent set is supposed to contain more than a small couple of individuals. Keeping up both, execution feasibility and egalitarian treatment of parents, is achieved by two measures, by (a) restricting the parent combination subsets to $2, 3, ..., N$-tuples with $N << \mu$, and by (b) dividing the set of parents into tiers and requiring the subsets to contain members from

different tiers.[52]

The cited articles describing scatter search are outstanding compared to the majority of EA literature because in agreement with the spirit of the quote above by Fred Glover they follow the maxim of not only explaining, but also backing with justifying arguments every single algorithm detail and every decision made by its designer. Scatter search stands out in the here presented row of EA examples as the least nature-inspired algorithm, exactly because its design is based upon these arguments and not on abstracting MNEA. As its concept is so different, the detailed comparison with other EAs is very instructive and expands the horizon of what shapes an EA can take. Concerning some of the algorithm ingredients, the justification arguments seem to make more sense in the context of discrete problems, so one can ask why it works also well in $\mathbb{R}^n$. In contradiction with reports documenting a competitive performance of scatter search [114, 212], it is drastically underrepresented in conference and journal publications on continuous-domain optimisation.

Algorithm outline

After having exemplarily described the spirit of scatter search, here comes the entirety of the abstract concept following the template given by Glover in [168]. To make the connection with the other EA pseudocodes, $\mathcal{P}_{\text{RefSet}}$ can be associated with the parent population $\mathcal{P}_{\text{F0}}$, $\mathcal{P}_{\text{new}}$ corresponds to the offspring $\mathcal{P}_{\text{F1}}$, and $\mathcal{P}_{\text{DivSet}}$ is a proto-population. $\mathcal{P}_{\text{RefSet}}$ is of fixed size μ and $\mathcal{P}_{\text{DivSet}}$ is of larger size N_{div}.

Algorithm 12: The scatter search algorithm

```
P_RefSet ← GenerateEmptyPopulation(μ)
P_DivSet ← GenerateDiverseSet(N_div)
P_DivSet ← ImproveMembersByLocalSearch(P_DivSet)
P_RefSet ← update(P_RefSet,P_DivSet)     // increase fitness & diversity
while ChangeDetection(P_RefSet) do
    L ← ListAllReasonableLengthSubsets(P_RefSet)
    L_NS ← OnlyNewSubsets(L)
    P_new ← RecombinationMethod(L_NS)
    P_new ← ImproveMembersByLocalSearch(P_new)
    P_RefSet ← update(P_RefSet,P_new)    // increase fitness & diversity
```

Thus, the size of the offspring population $\mathcal{P}_{\text{new}}$ is solely determined by how many subsets get generated after having imposed a reasonable N-tuple cutoff criterion and having filtered out subsets already seen. If there have been only few newcomers to $\mathcal{P}_{\text{RefSet}}$ during the latest iteration, then the number of unseen subsets will be comparably small, too. The recombination method must be able to handle parent

[52] In the case of $\mu = 4$ parents there are 10 possible subsets (namely: [1,2], [1,3], [1,4], [2,3], [2,4], [3,4], [1,2,3], [1,2,4], [1,3,4], [2,3,4]). But if the four parents are split up into two tiers (say [1,2] and [3,4]) and if we impose the constraint that subsets of interest need to contain at least one member from each tier, then the number of possible subsets reduces to 8 (because [1,2] and [3,4] are not allowed any more). For larger μ and if subset sizes close to μ are not allowed anyway, the reduction becomes more substantial.

sets of various sizes. Note that the employed break-off criterion is not so common in the world of EAs. But in case no random number needs to be generated inside the loop, rendering it wholly deterministic, this is the most obvious thing to do. Optionally, this loop can be nested into an external loop the following way:

1. perform the main loop until it stops,
2. reinitialise $\mathcal{P}_{\text{DivSet}}$,
3. conserve a certain elite among $\mathcal{P}_{\text{RefSet}}$, but diversify it using the new $\mathcal{P}_{\text{DivSet}}$,
4. repeat `max_iter` times.

The few seminal publications remain mostly on that abstract level in their description of scatter search, insisting that descending into the layer of code implementation might limit the broadness of the reader's perception of the concept, but also respecting that all usefulness comes from the particular implementation and implying the modest view that, given one sample algorithm, a better version might always just lie around the corner. But as the interest of this list of algorithms does not only lie on the conceptual ideas, but also on the inner workings of robust state-of-the-art optimisers, it will be tried to outline those particular subroutine implementations, which were proved competitive by Derrac et al. [114] by turning to the sources [172, 201, 249].

Particular method implementations

The routine creating $\mathcal{P}_{div}$ is called *diversity generation method*[53] by the authors and the particular implementation of [172, 201, 249] addresses bounded continuous problems and is based on dividing the bounded allowable range of each parameter into four equal segments. For each new solution candidate and each parameter there is a two-step process of first determining one of the four segments by random and then projecting a uniform random distribution onto the chosen interval to determine the final value. For every next member of $\mathcal{P}_{div}$ the four segments are represented with probabilities inversely proportional to their occurrences so far.

The choice of the local search (LS) technique, the *improvement method*, is completely free. Reported in [201] are the Nelder-Mead algorithm (also called downhill-simplex algorithm) [323] and the Solis and Wets algorithm[54] [424]. Hvattum et al. examined the performance of scatter search switching through eight different LS heuristics, six well-known LS techniques and two scatter search routines yielding in these cases two layers of nested scatter searches. It is common not to let the LS converge completely in order to limit cost. Then the main parameter to be set with any LS is the LS depth, i. e. how many function calls to allocate to the improvement of one new offspring. This choice mainly depends on the dimensionality n of the search space.

[53] The common literature names are printed in italics in this paragraph.

[54] It is basically the same as a (1+1)-ES with a mutation operator in form of a standard normal distribution and a simple step size adaptation rule: increase after several successes in a row and decrease after a series of failures.

Glover et al. call the parent population $\mathcal{P}_{\mathrm{F0}}$ the "reference set" or short *RefSet* and the rules for composing and modifying it the *RefSet update method*. They divide the *RefSet* into two tiers of sizes b_1 and b_2 (in [201] $b_1 = b_2 = 10$). Generally, elements gain access to the first tier (the elite) through a good value of the objective function, i.e. if one has a better value than the worst one occurring in the first tier, it will replace it. Access to the second tier is granted due to the property of enlarging the diversity of $\mathcal{P}_{\mathrm{RefSet}}$. The measure for diversity enhancement is an element's "closeness" to the elite cloud, or more precisely, the euclidean distance of genotype to the cloud's closest element, which is named d_{min}. Hence, a tested element with a d_{min}-value larger than the shortest d_{min} found among the second tier will replace that too close element. Applying the *RefSet update method* to initially create $\mathcal{P}_{\mathrm{RefSet}}$ means taking the first b_1 elements from the ordered $\mathcal{P}_{\mathrm{DivSet}}$ and filling the rest of empty seats in $\mathcal{P}_{\mathrm{RefSet}}$ up from the remaining members of $\mathcal{P}_{\mathrm{DivSet}}$ according to their d_{min}-values. Inside the SCS main loop the method's application means testing all newly created and improved elements first for incorporation into the elite and, if this fails, also for diversity-enhancement of $\mathcal{P}_{\mathrm{RefSet}}$. Interestingly however, the second entry gate seems to be closed in the main loop in [201] where SCS was adapted to real-parameter optimisation.

The *subset generation method* of [201] is restricted to all subsets of $\mathcal{P}_{\mathrm{RefSet}}$ of size 2 with at least one member that did not yet exist during the previous iteration of the main loop. In [172] also larger tuples are accounted for. Sets of triplets and quadruples are derived by adding the one or two best elements not formerly contained in a subset to it. Finally, sets of sizes $i = 5...b$, where b is the size of $\mathcal{P}_{\mathrm{RefSet}}$, are formed by joining the best i elements.

The recombination operator in a scatter search algorithm is called *combination method*. The combination method has to deal with any parent subset, also of sizes $N > 2$, generated by the subset generation method. It has to follow a systematic, desiredly deterministic set of rules to create new solution candidates. Random sampling in the parent neighbourhood is here not only considered unelegant, it is also unnecessary if the combination method is fed each subset only once. Figure T.18 illustrates how one can proceed in designing a generalised deterministic multi-parent combination method producing coordinate system-independent offspring. But against all dogmatism there are also approaches with a bit of randomness: Herrera et al. [201] in their study present a comparison between using the arithmetic mean and the BLX-α operator with $\alpha = 0.5$ on subsets only of size 2.

Discussion

At first glance scatter search presents itself as a much less nature-inspired EA than most other ones. The emphasis on avoiding a collapse of gene pool diversity and the excessive use of randomness hints rather towards inspiration from problems of other optimisation techniques. When a gradient method fails it can be seen as the consequence of assumptions oversimplifying the search task. When EAs fail on easy problems because one or a few mediocre genomes take over the whole population too early, then it can be seen as the consequence of oversimplifying assumptions on evolution, like 'let "selfish genes" compete and ever better solutions will emerge with

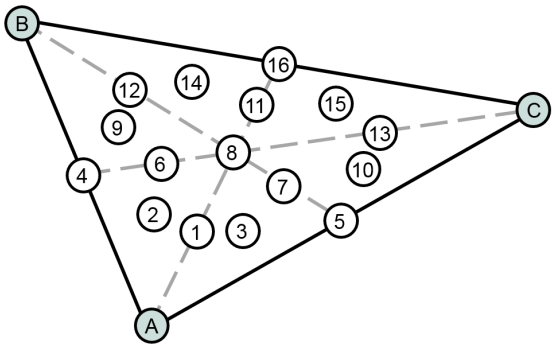


Figure T.18 Generalising deterministic multi-parent sampling.
The left diagram shows that two parents (the black dots) form a line, whereas the three parents on the right define a plane. In the case of two parents, with the help of random distributions one can sample segments of the line or also the space beside the line. But without random numbers it is not possible to sample the volume next to the line in a fair and neutral manner, coordinate system-independently. The idea behind Glover's quote demands to think of a sampling rule to create points on the line which will be beneficial to the algorithm. The deterministic offspring creation routine will be repeated exactly once with each pair of parents in $\mathcal{P}_{\text{RefSet}}$. With two parents $\vec{x}$ and $\vec{y}$ and $d = |\vec{y} - \vec{x}|$, what are the options? The most obvious offspring choice is the centre point $\vec{x} + \alpha(\vec{y} - \vec{x})$ where α is $\frac{1}{2}$. The next idea: two points with $\alpha = \frac{1}{3}$ and $\frac{2}{3}$; then $\alpha = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$; but how about diversity, the population should not just shrink within the initially covered subspace, therefore: $\alpha = -\beta, \frac{1}{2}, 1 + \beta$ with e. g. $\beta = 0.3$, or five points, and so on. The next question is what to do with three, four, five parents, how to generalise for N parents? Looking at the example with two parents it can be interpreted that the line had been cut by the parent dots into three segments, and that each segment has been sampled once. Instead of a line segmented by two points, imagine a plane segmented by lines that can be drawn in, around, and through a triangle of three points and these lines separating areas and other lines to be probed. This is what had been done on the right. In that diagram, secondary lines connecting one edge of the parent triangle with the centre of mass have been added. Thereafter, 16 new points can be identified being centres of line and area segments. But these lie all within the parent triangle, how about exploration? Note that with a similar set of rules the same structure can also be built if the three points with the crosses are the three parents. The triangle also allows the definition of pyramid tips outside the initial parent plane.

a bit of random mutation'. Scatter search can be seen as a "less lazy" approach. An algorithm can't be devised with only the things to avoid in mind. Instead of making abstractions from MNEA, scatter search is the result of a very constructivist approach, of the effort to come up with a maximally generic and robust metaheuristic.

So how well can the MH be adapted for real-parameter search? Comparisons of scatter search applied to hard optimisation problems in bounded continuous domains with other state-of-the-art EAs were published by Derrac et al. [114] in 2011 and Hvattum et al. [212] in 2013 and document a quite competitive performance. Considering these results, scatter search seems to be drastically underrepresented in journals and conferences on EAs (while [114, 212] both benchmarked SCS on the CEC-2005 testbed, no SCS was present at the CEC-2013 competition on real-parameter optimisation or at the BBOB competition at GECCO-2013). The reasons for this discrepancy are hard to guess. Maybe the conceptual foreignness with respect to more nature-inspired EAs plays a role. Some features of scatter search indeed seem counter-intuitive in the context of experiences with most other EAs, they will be discussed in the following.

Random – makeshift or virtue? Glover's argumentation has to be contrasted with the argument that randomisation can also be a safety measure (see figure T.5 on page 487) and a countermeasure against a collapse of search space (shifts out of

the orbit plane in PSO, see section T.4.7). But on the side of scatter search one can weigh in the two facts that (a) the geometrical patterns generating new points are ever changing because they depend exclusively on the geometry of the parent subset, and (b) the deterministically generated offspring represents just a starting point for the local search, i. e. the point will shift and settle down nearby, so the resulting point will not stay the one computed by the deterministic algorithm. And if the search space has collapsed, then scatter search starts over again with a new $\mathcal{P}_{\text{DivSet}}$.

The cost and effect of local search (LS): In an n-dimensional search space any local search should be allowed to run for several times n function evaluations if substantial progress towards the next local minimum is desired. But this means the cost of an EA will also be multiplied by several times n. This cost explosion is prohibitive for many EA applications. Alternatively, in order to stay at constant cost, the affordable number of generations has to be divided by that factor. But then many EAs will not be able to converge, or they lose their impact on the whole algorithm. Another problem can be a loss of gene pool diversity if after the LS many chromosomes of the population look the same. Adding local search claims a huge cost and may tip the exploration vs. exploitation balance of a given EA when introduced. The next option is to afford LS only to a few population members, either by choosing worthy chromosomes or by random selection. But this is very likely to destroy the global search power of an EA because the few improved chromosomes showing up early in the search will attract the focus of the EA even though they may be in a mediocre local minimum. Then the low number of local search spotlights stands against the high number of local optima and the curse of dimensionality is in the way. These can be reasons for an EA practitioner's weariness about LS because in practice bad experiences with it are just too easily available.

Here it must be noted that the conflict between LS and the host EA is damped in scatter search (a) because the feeding of similar chromosomes into the LS engine is systematically avoided (because there are no small mutations, and identical subsets are skipped), (b) $\mathcal{P}_{\text{RefSet}}$ is very small, (c) the algorithm can converge in very few iterations of the main loop, and (d) there is no local search unfairness, it is afforded to each new chromosome.

The small size of the population: $\mathcal{P}_{\text{RefSet}}$ must be small because all available subsets (up to a given n-tuple limit) will be tried out, and depending on the *combination method* the processing of one single subset leads to several solution trials. In [201] the $\mathcal{P}_{\text{RefSet}}$ consists of only 20 chromosomes in a setting where the algorithm would consume up to 5×10^5 function calls. EA practitioners must be weary about the intrinsic feasibility limit at such small population sizes because in EA applications one counts on the power of the large swarm to capture, or at least get along with the highly structured shape of the fitness function in n dimensions, and all this without much mathematical intelligence, only relying on the data in the chromosome library of the current population. But scatter search seems to work anyway. The many restarts of the main loop in connection with the diversity supporting mechanisms probably contribute to that.

The swarm dynamics in the box containing the search space: $\mathcal{P}_{\text{RefSet}}$ has two tiers of size 10 in [201]. By Glovers scatter search template [168], acceptance into the

first tier is by fitness, acceptance into the second one by diversity. In the context of chromosomes being binary strings it makes sense to have CO partners with maximum Hamming distance as it means maximising the number of loci where a CO operator can cause bit flips. But what happens in a continuous search space? Since 10 is such a small number, and entry into the first tier goes without diversity condition, it can be assumed that often the first tier will concentrate in one small subregion of the search space. If it is a corner then the diversity acceptance criterion of the second tier will lead to its members jumping ever farther away from that corner until they end up in other corners of the search space. As described earlier, many other EAs gain their efficiency in the fight against the curse of dimensionality and the unknown structure of the fitness function by shrinking steadily, by continuously abandoning more and more regions of the search space, by sinking down through the range of scales from large-scale to fine-scale structure at the end. With that background thought it is hard to imagine what the recombination between high-distance parent sets from the two tiers is good for. This corresponds to creating offspring along lines crossing the whole search space whereby only the offspring in between parents will be feasible and offspring outside the parent interval will be outside bounds. If the elite covers several spots forming a surface, then recombination among mixed-tier subsets means checking along lines and within surfaces perpendicular to the elite surface. Nothing wrong with that. But if the elite surface divides the search space into unequal parts, then all 2^{nd} tier members can be expected to show up only on one side of it. The question of clustering among the second tier is an aspect not addressed in the algorithm descriptions. Interestingly, in the continuous scatter search by Herrera et al. the two tiers exist only in the (re)initialisation phase when $\mathcal{P}_{\text{RefSet}}$ is generated from $\mathcal{P}_{\text{DivSet}}$. Later on in the main loop only the fitness acceptance criterion allows entering $\mathcal{P}_{\text{RefSet}}$ and the tier boundary is dissolved [201][55]. Thus, in their algorithm the second tier members do not get pushed further and further away from the elite, rather, $\mathcal{P}_{\text{DivSet}}$ and the second half of $\mathcal{P}_{\text{RefSet}}$, generated from $\mathcal{P}_{\text{DivSet}}$, are replaced every couple of main loop iterations. So one has to imagine that during several runs and restarts the elite gets new data input and probably shifts or jumps a couple of times. In the variant proposed by Hvattum et al. [212] acceptance of a solution into $\mathcal{P}_{\text{RefSet}}$ occurs under this condition: it is better than the current best $\vee$ ((it is better than the current worst) $\wedge$ (the closest point in $\mathcal{P}_{\text{RefSet}}$ is further away than a given limit d_{thresh})). The discrepancies in the diversity acceptance rules between [172], [201], and [212] highlight two aspects. According to Glover's quote, advantageous and disadvantageous decision rules have stronger effects the more deterministic the algorithm is. Secondly, rules that are suitable in a discrete problem might be a drawback in other contexts, this means the low-level routines have to be changed or adjusted when a metaheuristic is switched from context to context. But the NFL theorems have been telling that all along. The work by Hvattum et al. explores this aspect further by benchmarking the MH with many more LS variants than only Nelder-Mead and Solis-Wets. On the swarm dynamics of that part of $\mathcal{P}_{\text{RefSet}}$ which had only been selected for fitness, one can say that it exhibits the same ability as a CMA-ES population or a Nelder-Mead simplex, namely to be able to take on a very

[55] see pseudocode on page 455

squeezed shape able to follow down very narrow diagonal valleys.

Summary

Scatter search represents a wholly different approach to how to design an evolutionary algorithm. It seems to work well if it is done right, but doing it right takes a little more effort than with other EAs, and some of the effort will have to be repeated with any new optimisation task of different nature. Like in any EA with LS, the LS depth is a parameter with huge impact. In scatter search the decision on the LS depth has to be made in connection with the population size, the subset size limit, and the desired restart behaviour. The other sensitive points seem to be the RefSet acceptance rules and the choice of the LS routine. Scatter search could thus be problematic in cases where the topology of the training problem does not exactly match the features of an expensive real-world problem. Experimenting with scatter search is very useful because it challenges a look at EA details, and at the same time at the big picture, it shows that too much nature-inspiredness can act as a prison.

T.4.9 Hybrid EAs and memetic algorithms (MA)

Creating a *hybrid EA* can mean combining operators stemming from two or more different existing EA schemes in one single EA loop for offspring generation. In that sense the hybrid EA scheme outlined in chapter 3 represents a combination of ES and GA facets.

EC literature is often based on a broader definition where creating a hybrid EA means coupling an evolutionary algorithm with any other optimisation technique in any way. The definition is very broad because two optimisation algorithms can be nested, combined to work together with equal rights, made to take over intermittently, given a budget by a superior MH, just to name a few options. The combined optimisation techniques can be anything. If there is an EA in it somewhere, and if the EA part is not only at the low end of the hierarchy, then it is a hybrid EA. However, the coupling concept most often encountered is probably the one of an EA responsible for efficient global exploration calling up a local search (LS) subroutine for further optimising selected individuals. Scatter search, as explained above, is therefore a good example of that type of hybrid EA.

The term *memetic algorithm* is another one with a broad meaning. The neologism “meme” for the smallest unit, the atom, of cultural evolution was coined by evolution biologist Richard Dawkins [103] in search for an analogy to the word “gene”, the smallest meaningful unit of carried information in biological evolution. Cultural evolution happens when beings, which are intelligent and capable of learning, pass on their improved knowledge by communicating it to descendants and fellows. Chimpanzees using a grass blade for catching ants and passing down the technique by teaching and learning provide a good example of a hereditary meme. Examples of human cultural evolution through improved memes “going viral” are the techniques of using a saddle for riding a horse, making bronze, copper, iron, glass, “googling” something, citation styles in scientific literature. The memepool is not only shaped by population-based processes like differential reproduction rates

and random fluctuations, but also by the driving forces of curiosity, imitation, and purposeful thoughtful improvement. Mechanisms like invention, copying, idea diffusion, and schools of thought emerge. This yields a much faster evolution process, the human cultural evolution provides the best example. It took a few tens of thousands of years to evolve to the current state, whereas MNEA has been working on this planet for billions of years. The dynamics of meme populations develop on two levels with different speeds: following their carrier's population dynamic (e.g. pottery decoration styles of the Stone Age) and through quickly being communicated horizontally (e.g. viral videos, or scientific ideas like numerical methods).

Moscato [309] proposed the EC concept of *memetic algorithms (MA)* in order to allow a similar jump in quality. What a memetic concept should mean in the field of EAs is straightforward: the most useful knowledge or concept that a single member of a population can carry is the local search (LS) strategy. If the LS is to become a meme, then this implies the goals to make the individual's LS rules and strategy choices as flexible as possible and to allow the most suitable LS meme to evolve and "go viral" across the population. The EA designer's task is then to find the most efficient heuristics and code framework to let the memes evolve and adapt quickly to the problem and landscape currently explored by the population. Unfortunately, there is some confusion in EA literature and hybrid EAs with inflexible static local search subroutines are often labelled as MA, too.

The development of MAs seems to be most vivid in the field of discrete combinatorial problems, whereas constrained real-parameter optimisation seems to be a minor fraction. The journal "IEEE Transactions on Systems, Man, and Cybernetics" had a special issue on MA in 2007 [332], and "Soft Computing - A Fusion of Foundations, Methodologies and Applications" brought one in 2008 [333]. These and the thesis of N. Krasnogor [237] could be a starter for the reader interested in MA.

One local search requires substantially more function calls than the number of dimensions n that have to be explored. It is easy to make the calculation cost of an EA explode when systematically inserting LS subroutines without restricting their use carefully enough. The balancing of exploration versus exploitation grows to a severe and complex problem. Determining the exact guidelines for the 'where', 'when', 'to which individuals', 'how thoroughly', and 'what type of heuristic' should the subroutine be applied to becomes a very demanding task. Thus, MA comprises a multiplication of the fields of nonmemetic EA, local search strategies, control strategies, and adaptation schemes.

T.5 Multi-objective optimisation (MOO)

When Mother Nature's EA works on improving the genetic code for a mouse, the optimisation process is inherently multi-objective (MO). Some mice die because a bright fur colour makes them easy to be discovered by predators, some die because they are too slow when fleeing in danger, some die because their immune system cannot handle a disease. There are innumerable ways by which fate can shorten a mouse's life. This means constantly different selection pressures are acting on a population of mice in a randomised way, and a well-designed mouse is the best

compromise serving all the needs of being fast, attentive, energy-efficient, fertile etc. while not being overly prone to any of the common causes of death. Chipmunks, squirrels, rats, mice, rabbits, they all have a common ancestor, but because of gene drift and population separation over long time they have now evolved into different species and represent slightly different compromise solutions to the multi-objective challenge scenario posed by their environment, one speaks of different *ecological niches*.

It would be easy to implement similar EA schemes of multiple pressures (i. e. objectives) acting in a randomised manner. But normally when resorting to multi-objective optimisation algorithms, the goal is not that evolving populations drift into different corners from run to run illustrating everytime a different way to compromise. Many runs would be needed to get a feeling of the ensemble of possible good compromises. Generally, the intention behind the use of an MO algorithm is to learn about the spatial distribution of good compromises in the search space and their distinction against inferior compromises. This can be discussed in the framework of *Pareto-optimisation*[56].

Pareto-comparability

Instead of one single objective function f_{obj} we are dealing now with two or more scalar functions $f_1(\vec{x}), f_2(\vec{x}), ..., f_k(\vec{x})$ to be minimised. A solution is said to be Pareto-inferior to others, if there is a solution in the population being better with respect to at least one objective and worse in none. A solution is called Pareto-efficient if it is Pareto-superior over other points and there are no points Pareto-superior to it. All Pareto-efficient solutions together form the *Pareto-front*. In order to optimise one objective going out from a Pareto-efficient solution, one has to accept losses in at least one other objective.

A Pareto-optimiser has two functions: firstly, it should give an answer to the question whether the defined objective functions create a goal conflict, and secondly, if there is a goal conflict, then the optimiser should improve the population of solutions such that the structure of the Pareto-front is revealed and that it closely approximates the true limits of feasibility.

Taking the example of aeroplane design, two exemplary goals not creating any conflict would be the minimisation of structural weight (for a given outer shape) to serve on the one hand the purpose of minimising fuel consumption and on the other hand the maximisation of the climbing rate. An exemplary combination of objectives giving rise to goal conflict would be the maximisation of mechanical stability, which tends to increase material usage, versus weight reduction. Another example is the wing profile. Every aeroplane designer has to make a decision on the balance between thin delta wings allowing high speeds or long and bellied sailplane wings achieving the required lift in the most economic way. A multi-objective algorithm is used when support is needed for making such kinds of decisions. A multi-objective EA has to move away from search space domains representing Pareto-inferior solutions, it has to push the population's Pareto front ahead, and it has to create a dense distribution within the Pareto-front in order to make its structure visible.

[56]The name comes from the Italian engineer and economist Vilfredo Pareto (1848–1923).

Example of a MOEA: NSGA-II

One of the most popular MO-EAs was presented by Deb et al. [108] in the form of the NSGA-II algorithm, a GA made capable of coping with MO problems by a subroutine called *nondominated sorting (NS)*. The GA enacts its selection pressure purely based on the individuals' ranking in the population (unlike fitness-proportional selection in traditional GAs). The specialty of the algorithm is how the population members are ranked in each generation. The sorting procedure starts with the distinction between those solutions forming the Pareto front and the other solutions. The Pareto front members are assigned the first nondomination rank. Next, the Pareto front members are taken away and it is asked which elements would now form the Pareto front. These elements are assigned the second nondomination rank. The procedure continues until all members of the population are assigned such a nondomination rank. The sorting into these Pareto front ranks is the primary sorting mechanism. The secondary sorting mechanism does not intermix these groups any more, it only re-arranges the elements within each group. The measure used for re-arranging is a measure of the crowding within the Pareto front going as a staircase or bending line through the f_1-f_2-space (or warped as surface into the f_k-space in general): along each f_k-axis, one dimension after the other, the two distances to the nearest neighbours on both sides are collected, along each direction there is a normalisation taking into account the extension of the whole set, and the average of all these normalised next-neighbour distance measures is called the crowding distance. The two solutions at the extremal positions along each direction obtain an additional fitness offset. This setup is intended to gear the NSGA-II to create more new candidate solutions along sparsely populated regions of the Pareto-front than within areas which have already been densely scanned.

Of course, many other EAs can also be based upon this sorting procedure to turn them into MOEAs. Note that the secondary sorting procedure exhibits an inherent sensitivity towards the scalings of the objective functions. For instance, re-formulating one of the objectives on a logarithmic scale will cause the algorithm to behave differently.

T.6 Summary on the EA overview

Few people have a working knowledge about quantum mechanics, but many have an idea of how evolution works. The concepts of mutation, variation, information recombination, selection pressure, and differential reproduction rates are simple. Putting these ingredients together to make a search engine working in the blind search manner is not always the most elegant or efficient way to tackle real-world optimisation problems, but it often works. Evolutionary algorithms can be made quite robust against being trapped in local optima, their applicability relies on few assumptions about the search space. This allowed EAs to become popular and a common global search method, and it renders them worth a consideration when confronted with many real-world optimisation problems.

It is not trivial and not always possible to tune a new EA idea built from scratch to competitiveness among state-of-the-art optimisers. But it is pretty easy to set up

an EA routine that works. The nice thing is that in many cases the interfacing of the EA with the optimisation problem is fairly straightforward because the optimisation problem is treated in a black-box manner. A few educated guesses on the type of most suitable EA approach and the main implementation settings can prevent many negative user experiences (and the present text is intended to enable these educated guesses for readers of any background). Therefore, the decision to use an EA often does not hinge on a narrow estimation of the pure computational cost, but on a somewhat broader comparison of practical scenarios: How does the added computational cost of using an EA ($\sim 10 - 10^3$ times more function calls than deterministic methods) and the necessary implementation work compare to the cost of one or more engineers working through the mathematical theory necessary to devise and implement a more targeted and suitable alternative optimisation framework? What is the risk that the initially envisaged mathematical theory turns out not to be general enough? When future modifications to the project will affect the optimisation problem, how does the effort to modify the EA treatment compare to the effort to update the specialised deterministic procedure? Will the alternative mathematical treatment be as flexible to handle eventual extensions of the optimisation problem as the EA treatment will be? What will the maintenance cost of the to-be-programmed optimisation framework be (potentially including handing the method down)?

EAs can have issues. If chosen and set up wrongly, they can be wasteful, stagnate in diagonal valleys, be out of balance with respect to local vs. global search, and suffer loss of gene pool diversity. EAs are simulations of populations of vectors in a search space, simulations of a swarm of particles jumping or gliding through a potential landscape. The simulated population systems exhibit their own motion laws. Therefore, of course, wrong settings are possible where the motion laws are dominant and the landscape to be searched is not, or where the swarm dynamics interact undesiredly with landscape features so that its motion pattern is not suitable for global search. But since EAs are composed of conceptionally simple ingredients, these failures can be detected, analysed, understood, and counteracted fairly easily too by tracking the statistics of populations, and by modifying the ingredients of the EA. Often, these occasions offer a possibility to learn about the nature of the optimised problem itself. The latter stance is of course true for any kind of optimiser, not just an EA. But the swarm dynamics of EAs can often be described in easy and intuitive terms, and this will be reflected in the lessons learnt about the search landscape. By contrast, when learning from the failure of a very intricate method being devised and described itself in very complex mathematical terms, the undesirable landscape features discovered upon the method's failure may also reflect that same complexity level when being described. Think of transitioning from a smooth to a noisy function. This will bust many definitions making mathematical analysis easy, but in the efficiency, applicability, and the underlying ensemble of assumptions of most EAs this transition will make no dent. In an EA framework one will still be able to discuss average local gradients and distributions of local optima. It is indisputable that the best solution offering the broadest learning potential generally lies in an analytical treatment of the physical and mathematical properties of a real-world problem allowing to deduce a dedicated and reliable global optimisation routine on that formulation. But as soon as generic optimisers are considered, EAs

should definitely be on the list.

Mother Nature uses for her different evolutionary algorithms variable sets of ingredients (haploidicy, diploidicy, chromosomes, CO, introns, exons, HGT, DNA repair mechanisms) and the successful ones of her EA trials still exist today in bacteria, green algae, vertebrates, etc.. Historically, the field of EAs (or broader: *evolutionary computation, EC*) was subdivided into different schools of thought (e. g. ES, EP, GA, GP, SA), but as more and more intermediate concepts and hybrid EAs are published, the scope of evolutionary computation is getting broader and more unified. The example of the CO operator, seen as a low-level heuristic called by a metaheuristic, and the fact that some form of it is used in a variety of popular EAs, some of them formerly seen as originating from supposedly different schools of EA thinking, and that there is one basic idea, but complete freedom of implementation of variants of the basic concept, underscores the validity of the view that EC should be seen as a field building on the usage of one diverse, comprehensive, and infinite toolbox of basic algorithmic elements.

The vast, comprehensive, but easily accessible collection of EC concepts invites for do-it-yourself approaches as well as for transferring established elaborate EAs to new applications. It invites for experimentation on the side of the optimiser as well as for tinkering with the formulation and representation of its target. Hard real-world problems challenge programmers to improve the optimisation algorithms while testing them on hard and telling test functions, and on the other hand the trials with applying different optimisers on a problem of real-world interest motivates users to rethink and improve their black-box problem formulations and to learn about the problems' nature.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
a_i	lower bound of search space along i^{th} dimension
b_i	tier sizes in scatter search description
b_i	upper bound of search space along i^{th} dimension
C	covariance matrix
CR	crossover rate control parameter in DE description
c_1, c_2	attractor force scaling factors in PSO description
D	diagonal matrix
d	distance
E	energy
$\hat{e}_i$	unit vector
F	translation scaling factor in differential evolution
f_{obj}	objective function
f, g, h	generic functions
G	maximum number of generations (algorithm control)
g	generation counter
I	identity matrix

K_j	attractor location (i. e. chromosome memory) in PSO description
L	population subset list in scatter search description
N	population size
$\mathbb{N}$	natural numbers
$\mathcal{N}$	normal distribution
n	search space dimension
P	probability
$\mathcal{P}$	chromosome population
$\mathcal{P}_{\mathrm{F0}}, \mathcal{P}_{\mathrm{F1}}$	parent, offspring population
`p`	chromosome ("particle") in PSO description
`p.L`,`p.v`,`p.x`	particle local neighbourhood, velocity, position in PSO description
$\mathbb{R}$	real numbers
r	random number
s	mutation step adaptation factor
T	temperature
$\mathcal{U}, \vec{\mathcal{U}}$	uniform distribution (scalar, vector)
$\vec{u}$	offspring chromosome in DE description
$\vec{v}$	intermediate (translated) chromosome in DE description
$\vec{v}$	velocity in PSO description
$\mathcal{X}$	function domain
x	spatial coordinate, design variable
$\{\vec{x}\}$	set of chromosomes
$\vec{x}, \vec{y}$	chromosome, i. e. point in search space
$\mathcal{Y}$	function co-domain

List of Greek quantity symbols

Symbol	**Description**
α	recombination operator extension range beyond better parent
α	inertia parameter in PSO description
β	recombination operator extension range beyond worse parent
β	attractor force scaling factor in PSO description
γ	maximum number of generations (algorithm control)
$\delta, \vec{\delta}$	mutation step
$\vec{\zeta}$	chromosome, i. e. point in search space
λ	size of offspring population
μ	size of parent population (selected for "survival" and reproduction)
$\mu, \vec{\mu}$	mean value, mean vector
$\vec{\xi}$	chromosome, i. e. point in search space
σ	mutation step size control parameter
ψ	attractor force scaling factor in PSO description
ω	angular frequency

List of abbreviations

Abbreviation	Description
ACO	ant colony optimisation
ANN	artificial neural network
BBOB	Black-Box Optimization Benchmarking
BC	binary-coded
BCGA	binary-coded genetic algorithm
BCO	bee colony optimisation
BGA	breeder genetic algorithm
BLX	blend crossover
CEC	IEEE Congress on Evolutionary Computation
CMA-ES	evolution strategy with covariance matrix adaptation
CO	crossing-over, crossover
DE	differential evolution
DNA	deoxyribonucleic acid
DOE	design of experiment
DX	discrete crossover
EA	evolutionary algorithm
EAO	evolutionary algorithm optimisation (meaning optimisation by evolutionary algorithm)
EC	evolutionary computation
ES	evolution strategy
EP	evolutionary programming
FIPS	fully informed particle swarm
GA	genetic algorithm
GECCO	Genetic and Evolutionary Computation Conference
GP	genetic programming
HGT	horizontal gene transfer
IEEE	Institute of Electrical and Electronics Engineers
ILLIAC	Illinois Automatic Computer (historic computer family)
JSSP	job shop scheduling problem
KL	Karhunen-Loève
LCS	learning classifier systems
LS	local search
MA	memetic algorithm
MH	metaheuristic
MNEA	Mother Nature's evolutionary algorithm
MO,MOO	multi-objective (optimisation)
MOEA	multi-objective evolutionary algorithm
MVN	multivariate normal (distribution)
NFL,NFLT	no free lunch (theorem)
NN	neural network
NP	nondeterministic polynomial
NS	nondominated sorting
NSGA	nondominated sorting genetic algorithm
PC,PCA	principal component (analysis)

PDF	probability density function
PE	punctuated equilibrium
PSO	particle swarm optimisation
RC	real-coded
RCGA	real-coded genetic algorithm
RNA	ribonucleic acid
ROM	reduced-order model
SA	simulated annealing
SCS	scatter search
TSP	travelling salesman problem
UNDX	unimodal normal distribution crossover
VRP	vehicle routing problem
WHX	Wright's heuristic crossover

Appendix U

Evolution: some subtle facts and interpretations

U.1 Why death?

Why is death, if it becomes systematically established as part of the EA of real life, a competition advantage? An individual bacterium does not necessarily have to die; all existing bacteria have lived ever since the beginning of bacterial life. Concerning our death, the empirical scientist can just make the laconic remark, that if there were individuals programmed to live eternally, being equipped with an optimal repair mechanism for body cells, they would hamper the evolutionary progress (i. e. speed of adaptation) of their own species, so there aren't. The two reasons are the following: with being evaluated in terms of fitness and having produced offspring or not, the individual's life's purpose is fulfilled, the gene pool of the population in the next generation is optimised, and the quicker the generational cycle is completed the better. The speed of the EA iteration cycle is a competitive advantage by itself. Repeated genetic input from the past would be a drawback more often than not, and too old members of the population are draining the limited environmental resources. As concerns mutations of the gamete cells, as more mutations accumulate over time, there would be ever less connection between the genotypic information transferred to the offspring and the phenotype of the parent under fitness evaluation. An extended individual existence as senior is only of advantage if cultural knowledge has to be conserved and passed on. E.g. in a band of elephants travelling through the Namibian desert there have to be seniors knowing about water holes revisited only occasionally after many years.

The species of green algae represent many interesting intermediate steps between simpler unicellular life forms without sexual reproduction and death and other eukaryotic organisms with the two features. Some species of green algae[1] are in the lucky position of enjoying the advantages of recombination without the disadvantages of programmed death. These algae live normally as haploid cells and reproduce asexually by cell division. But there is also a sexual reproduction cycle where the cells of two algae fibres undergo pairwise conjugation. Connection tunnels are built

[1] e. g. Spirogyra: `http://en.wikipedia.org/wiki/Spirogyra`

between two cells from different fibres and the cells from one fibre migrate through the tunnels and form diploid zygotes. None of the cells die. The empty cellulose hulls remaining from one of the two fibres are the only pieces of dead matter left behind in the process. The zygotes divide up again into haploid cells, the default form in this case.

U.2 Bacterial conjugation

"Of course" some form of reconnecting lineage branches has also emerged among bacteria: there is a gene called F-factor ("F" for fertility) present in the genome of some bacteria throughout many species. This genetic text contains the code of proteins necessary to build a tunnel leading outside the bacterium equipped with a tip able to poke into other bacteria that happen to be nearby. Once such a tunnel connection exists, the F-factor gene sequence transfers itself into the poked bacterium and behind the transferred F-factor code still follows a gene sequence of variable length composed of whatever happened to be behind the F-factor when it was part of the genome of the old host. Note that the F-factor spreading thus in a population of bacteria matches almost exactly the definition of a virus, but a virus of invaluable benefit to the speed of evolution of bacteria. Recombining genetic material by other means than sexual reproduction is termed *horizontal gene transfer (HGT)*. For further information see [271], `http://en.wikipedia.org/wiki/Bacterial_conjugation`, and `http://en.wikipedia.org/wiki/Horizontal_gene_transfer`.

Appendix V

Test functions used for EA benchmarking

V.1 Test function from literature

The shifted rotated Weierstrass function

The shifted rotated Weierstrass function is the function F_{11} of the CEC-2005 test function suite [444]. This collection was compiled by Suganthan et al. for the competition on real-parameter optimisation hosted by the CEC-2005 conference. The function F_{11} is given by

$$\begin{aligned} F_{11}(\vec{x}) &= F'_{11}(\vec{z}(\vec{x})) \\ &= \sum_{i=1}^{n}\left(\sum_{k=0}^{K}[a^k \cos(2\pi b^k(z_i+\tfrac{1}{2}))]\right) - n\sum_{k=0}^{K}[a^k\cos(2\pi b^k\cdot\tfrac{1}{2})] + \Delta \end{aligned}$$

with $a = \frac{1}{2}$, $b = 3$, and $K = 20$. The suggested search domain boundaries are $-0.5 < x_i < 0.5$. $\Delta = 90$ is just a scalar offset; it has the purpose to shift the level of the global minimum away from zero.[1] Additionally, most of the CEC-2005 test functions are shifted and rotated. This means that the point $\vec{x}$ doesn't enter the function directly, but is transformed beforehand via

$$\vec{z} = (\vec{x} - \vec{\Omega}) \cdot \mathbf{M}.$$

The vector $\vec{\Omega}$ now indicates the new position of the global minimum, which would lie at $\vec{0}$ in the untransformed function. The rotation matrix $\mathbf{M}$ can also include stretching. In the case of F_{11} it has the condition number 5. The data for all

[1] In historic GAs the selection pressure is directly related to the fitness. Imagine the situation when the weight of a construction has to be minimised: once the early phase of optimisation is over and the candidate solutions differ less and less in their weight, such an EA will become less and less efficient because the differences in relative selection pressure vanish. There is always the option to offset the fitness function, whereby the ideal offset would be the weight of the globally optimal construction. But in real-world scenarios the value of the objective function at the global optimum is not known in advance. So, in the real-world scenario it is desired that the EA performance cannot be compromised too much by fitness scaling issues. Using a test function suite with different offsets can be seen as a safety measure against developing EAs with sensitivities towards fitness scaling.

the shift vectors and rotation matrices can be downloaded from P. N. Suganthan's homepage [443]. His site also hosts the Matlab® code for the test functions. A few of the CEC-2005 test functions, rewritten in Python, can be found here [430].

The shifted expanded Rosenbrock plus Griewank function

is the function F_{13} of the CEC-2005 test function suite. It is a so-called expanded function, which means that it is a 2D function expanded to cover an n-dimensional space according to

$$eF(\vec{x}) = eF(x_1, x_2, ..., x_n) = F(x_1, x_2) + F(x_2, x_3) + ... + F(x_{n-1}, x_n) + F(x_n, x_1).$$

eF8F2 is the expansion of two nested functions

$$F8F2(x_i, x_j) = F8(F2(x_i, x_j)),$$

and these two core functions[2] are the Rosenbrock function

$$F2(x, y) = 100(x^2 - y)^2 + (x - 1)^2$$

and the Griewank function

$$F8(x) = \frac{1}{4000}x^2 - \cos x + 1.$$

In the case of F_{13} there is no rotation matrix involved, just a coordinate system shift

$$\vec{z} = \vec{x} - \vec{\Omega} + 1$$

and the shift Δ of the function values is $\Delta = -130$. The suggested search domain boundaries are $-3 < x_i < 1$.

The FM-synthesis problem

The FM-synthesis problem of the CEC-2011 collection of real-world problems [102] is a function fitting problem where the target function is a signal of the form

$$s(t) = A_1 \sin\Big(\omega_1 t + A_2 \sin\left(\omega_2 t + A_3 \sin(\omega_3 t)\right)\Big)$$

which involves the principle of *frequency modulation (FM)*. The task is to find the parameter set $\{A_1, \omega_1, A_2, \omega_2, A_3, \omega_3\}$ that is able to reproduce the original signal, i.e. that minimises the area between the original and the trial signal to zero. In order to make the test function computationally slim, a discretised version is considered. The area integration turns into the summation of pointwise distances, or simply square distances. The specific problem instance described in [102] looks like this: the signal

$$s(t) = A_1 \sin\Big(\omega_1 t\theta + A_2 \sin\left(\omega_2 t\theta + A_3 \sin(\omega_3 t\theta)\right)\Big)$$

[2]Their names "F8" and "F2" stem from historic test function collections.

with the parameters

$$\begin{aligned} A_1 &= 1.0, & \omega_1 &= 5.0, \\ A_1 &= 1.5, & \omega_1 &= 4.8, \\ A_1 &= 2.0, & \omega_1 &= 4.9, \end{aligned}$$

and

$$\theta = \frac{2\pi}{100}$$

is defined on the grid

$$t \in [1, 2, ..., 100].$$

The objective function is the sum over the square offsets between trial and target function on that grid:

$$f(\vec{x}) = f(A_1, \omega_1, A_2, \omega_2, A_3, \omega_3) = \sum_{t=0}^{100} \left(s_{\text{trial}}(t) - s_{\text{target}}(t)\right)^2.$$

The search space boundaries suggested in [102] are $\vec{x} \in [-6.4, 6.35]^6$.

Anybody who has ever played with analogue synthesisers for producing FM sounds knows that FM is a funny thing. There are states where the resonance frequency dials of particular oscillators become extremely sensitive and twisting the dial by a tiny bit creates a totally different sound. Many different sounds are possible, and with more than two or three oscillators it becomes impossible to remember or get a feeling of what setting leads to what sound. On the other hand, a pair of oscillators $\sin(\omega_1 t + \sin(\omega_2 t))$ can be brought into an easily understandable and predictable state if $\omega_2 \ll \omega_1$, then it is nothing else but a siren. This has two important implications: (a) that the exact boundaries of the search domain are a crucial part of the problem description, and (b) that the search domain can be extended into regions where the ruggedness of the objective function becomes extremely high and there is no strong causality in Rechenberg's sense [379] (which would require that infinitesimal variations in $\vec{x}$ lead to infinitesimal changes in f_{obj}). Another aspect, in which the FM-synthesis problem is special, is that the impact of variations of the amplitudes has a different nature than the impact of variations of the frequencies. And the amplitude of the outermost harmonic function has of course the weakest impact. If these special properties are kept in mind and care is taken that a new problem instance is not made too easy nor too difficult, and therefore useless, then it is a substantial enrichment for each collection of test functions. The freedom to steer the window formed by the search space boundaries towards easier or more difficult areas or to span transition regions is in any case something very positive for the EA experimentalist.

The function F101 by Whitley et al.

The function **F101** by Whitley et al. [508, 511] is an expanded function. The primitive 2D function

$$F101(x, y) = x \sin\left(\sqrt{|x - (y + 47)|}\right) - (y + 47) \sin\left(\sqrt{|y + 47 + x/2|}\right)$$

has a unique global minimum on the diagonal. It is expanded according to

$$\begin{aligned} eF(\vec{x}) &= eF(x_1, x_2, ..., x_n) = F(x_1, x_2) + F(x_2, x_3) + ... + F(x_{n-1}, x_n) \\ &= \sum_{i=1}^{n-1} F(x_i, x_{i+1}), \end{aligned}$$

this time without cycling through from n to 1. As suggested in [511], the search domain boundaries were defined by $x_i \in [-512, 511]$.

V.2 A newly developed test function: the charged marble problem

Test functions are a tool for learning about search and optimisation algorithms. The tool has to be used by prospective EA users as well as EA inventors. The communities of researchers working on deterministic and stochastic search algorithms have created comprehensive collections of telling test problems with manifold characteristics and difficulty levels. Competitions are held among algorithm developers, whereby each competition is another occasion to compile a new library of test functions. Of course there has never been a real need to build a new class of test problems with special properties in the context of making the current EA choice. But during the process of familiarisation and experimentation with existing EA concepts and test problems, a very simple popular type of test problem with the added benefit of easy solution visualisation has evolved into a more challenging test problem while conserving the benefit of visualisability. The result is a test problem which is at the same time computationally simple, a challenging search problem, and one where solutions still can be visualised in such a way that humans can interpret solutions and judge on quality differences within fractions of a second. This can be identified as a class of test problems filling a gap. That this gap is otherwise populated so thinly, is somehow surprising. From the perspective of the EA developer, it is felt by the author as a big advantage during writing EA code, which is a continuous chain of decision-making, to be able to ask for a quick and intuitively graspable feedback about changes in the algorithm performance. Therefore, the test problem class was published at a conference on EC [428], and a short description will follow below.

The benefits of an intuitively understandable visualisation

Sound evaluations of algorithm performances are of course only possible through large statistics on many test problems. The larger the statistics, the finer differences can be revealed. But in the early stage of developing program code for trying out a new algorithm idea, when many fundamental decisions with large effect have to be made, it is very advantageous if quick experiments are possible with a feedback loop of seconds so that a trial-and-error scheme can replace many arbitrary guesses and speculations. One option is to use a computationally very simple test function where a small histogram of optimisation results can be produced quickly, and to examine how the histogram shifts when the setup of the algorithm is changed. Another option is to look at single optimisation histories and analyse particular chromosome

modification steps and their effect on the search. Both approaches are necessary and complementary ingredients to the process of brewing new optimisation algorithms.

The analysis of single optimisation histories is substantially easier if a simple way of visualising solutions exists that can be easily interpreted instead of a row of meaningless numbers. For the Rastrigin function [444]

$$f_{\mathrm{Ra}}(\vec{x}) = \sum_{i=1}^{N} \left({x_i}^2 - \alpha \cos(2\pi x_i) + \alpha \right) \quad \text{with} \quad \alpha = 10,$$

the visualisation scheme depicted in figure V.1 can be taken as an example. The analysis of optimisation histories of such a visualisation does not only reveal how quickly the algorithm improves the solution but also through which intermediate steps the search proceeds. Single mutation or recombination steps can also be analysed by comparing the plots of the individual solutions. This offers a creative programming practice of implementing a new EA idea, watching a handful of optimisation runs, changing an implementation detail, watching again, and so on. This is very nice, except that it is useless unless the examined test problem is challenging. Pushing oneself to develop code for an EA performing better and better on an easy test problem unavoidably leads to disappointment at the end of the day when the supposedly improved code is tested again with large statistics on the other test problems. Some non-challenging test functions (the Rastrigin function, or others, e. g. parabolic potential, sharp ridgeline, Rosenbrock function) have their justification in the development process due to their distinguished properties checking the optimiser for particular flaws. But would it not be nice to have a test function where one is able to combine and scale several of those properties that lead to truly challenging optimisation problems, but which is at the same time visualisable in such a simple manner that the programmer can judge at a glance whether her or his work of algorithm architecture develops in a fruitful direction or not?

Taking another simple test problem ...

Inter-particle potential problems are a popular class of test problems for EAs because they are able to feature high numbers of local optima. E. g. problem Nº 2 of the CEC-2011 competition on optimising real-world problems is the search for Lennard-Jones cluster structures yielding the minimal amount of potential energy of N particles within each other's Lennard-Jones potentials [102]. Taking the coordinates x_i, y_i, z_i of the N particles as the parameters to tune, how does the objective landscape look like? That the local energy minima must be separated by energy barriers is easily understandable. In order to go from one settled crystal structure to an alternative one, energy has to be invested first. The energy cost is needed to either break single particles away from their neighbours and carry them from one point on the cluster surface to a different one, or for shifting crystal layers against each other from match point to match point, or for squeezing single particles through layers of other particles. Other variations of the problem are to minimise the potential energy of N electrons with possible locations confined to a given volume or surface. Among the simplest instances of that problem class are the tasks of minimising the

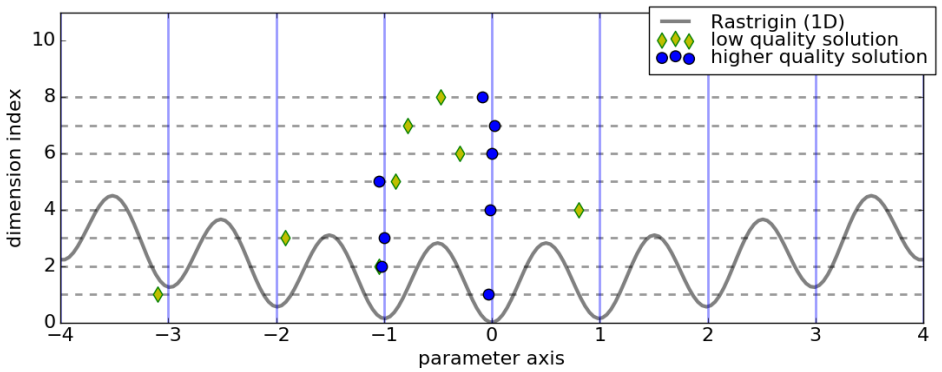


Figure V.1 A visualisation for solutions of the Rastrigin problem
The Rastrigin function is a sum of contributions from each entry x_i of the chromosome. Thus it is a separable problem and the parameters can be tuned one after the other. In the visualisation above, showing two solutions of the 8-dimensional problem instance, each parameter x_i is a dot that can move horizontally and which contributes to the sum according to the potential shown in grey. Therefore, each dot would have to be pushed to the centre position in order to realise the one solution representing the global minimum. Interestingly, watching EA optimisation histories (i. e. snapshot sequences of the current best of each generation), one can clearly distinguish a traditional GA from a traditional ES. The GA, due to the recombination operator, tends to replace only parts of the best current chromosome at a time. By contrast, the ES lets the whole picture flip if there is a change. A conventional ES pushes each dot towards the centre of its local valley and only occasionally a dot jumps into a neighbour valley, whereas in the optimisation histories created by a GA the dots are very likely to jump through a few valleys before they settle in one of the better ones. It looks like each dot conducts its jumping search independently, and this is indeed what a conventional GA does. It tries recombinations of mutated partial solutions of the separable problem. Therefore, scrolling through two or three optimisation histories does not only indicate how quickly the algorithm improves the solution quality, it can also reveal useful information about how the algorithm functions. Unfortunately, due to being separable, the Rastrigin problem is too simple and not suitable to judge EAs.

potential energy of N each other repelling particles free to move on a linear track, on a circle track, or on a sphere surface.

... and making it harder ...

The case of minimising the potential energy of N charged particles sliding on a circle line was taken as the starting point. This can be trivially visualised and it is depicted in figure V.2. It is very easy for a person to decide which one of two solutions is the better one. Evolution gave us a good eye for evenness, and of two competing trial solutions we will in many cases be able to tell at a glance which is the more even distribution. Secondly, everybody knows about forces proportional to $1/r^2$ (potential $\propto 1/r$) through haptic experience with magnets. Therefore, we have a feeling for the work necessary to deform the case of ideal even order and get to the one we have to judge. But of course, the task of finding the global energy minimum poses no real problem even for simple algorithms. At any moment each particle is pushed away from its closer neighbour and towards the farther one which explains why the simplest evolution strategy, downhill-simplex search, simulated annealing, or a gradient follower can solve the problem within few iterations. The equidistant distribution is globally optimal, and there are infinitely many optimal solutions. They can be reached through rotation of the system or swapping of particles.

However, if a second type of potential energy is introduced, a hilly track potential,

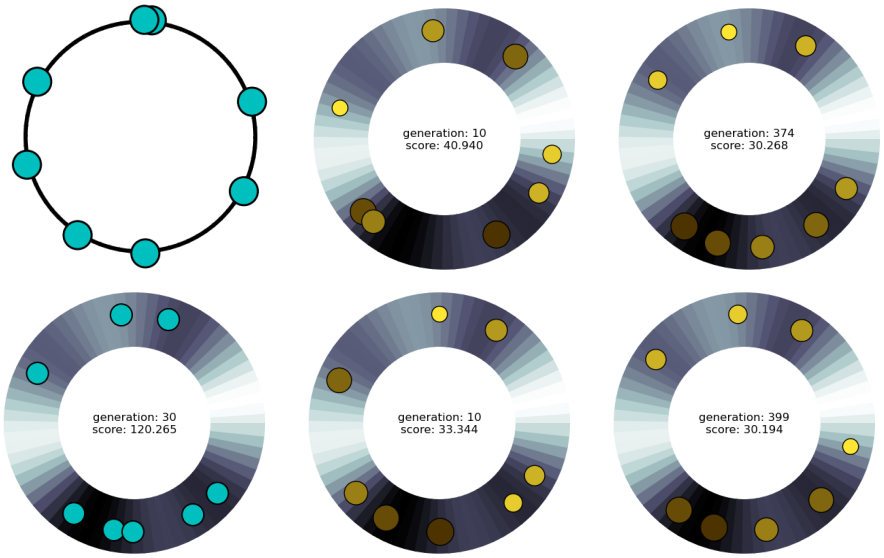


Figure V.2 Adding difficulty to a simple particle potential problem
The picture in the top left corner shows a visualisation of the simple test problem of repulsive particles on a ring track. Humans can interpret the solution quality quickly because we can look at the distribution of dots in terms of evenness. In most cases it takes just an instant to decide which one of two competing solutions is better. Through play with magnets we have an intuition for forces diverging at close distance. Asked to judge according to a criterion of potential energy, we can infer from the two overlapping dots on the ring that the quality of this solution must be particularly poor. The picture in the lower left corner corresponds to the first step of making the problem more difficult by the addition of a track potential. A different quality of hilliness is thus introduced into the search space because now there are barriers separating local energy minima. Another measure increasing the search difficulty level drastically is to make the particles nonidentical by giving them different weights. This adds a combinatorial aspect. Solution visualisations of this problem instance are depicted in the remaining four pictures. The two plots in the middle show solutions of low and intermediate quality, due to too close neighbours, heavy marbles in elevated valleys, and light marbles wasting space in deep valleys. The two solutions depicted on the right are near-optimal.

and by making the objective function the sum of both types of energies, the situation changes drastically. Now there is just one combination of particle coordinates left to represent the global energy minimum. Only those solutions reachable by particle swaps are still equivalent. Most importantly, there are now different local optima depending on the particles' distribution among the four available valleys. The local minima of $f(x)$ in the search space lie in energy valleys which are separated by energy barriers reflecting the fact that particles have to be pushed over hills on the track in order to roll down into a different valley like a marble. It has to be noted that the visualisability and the ease of picture interpretation are still given. This is intended to be shown by the second image in the first column of figure V.2.

... and harder

With the simple measure of making the particles non-identical by giving them individual weights, the problem can still be made much harder. How this can be easily reflected in the visualisations is shown in the remaining four pictures of figure V.2. Now it matters into which valley which marble falls. Near-optimal solutions are only found if deep valleys are not wasted with light marbles. With the identical particles, as soon as each valley contains one or two of them, a local minimisation routine will

succeed in finding one of the few near-optimal solutions. But with the non-identical marbles of different weight, a combinatorial aspect enters the problem, and due to the many possible permutations, the number of high-quality solutions multiplies, and they are all distinguishable.

It turns out that the problem is so challenging, that neither state-of-the-art EAs like CMA-ES, DE, and PSO, nor the in-house EA, can exhaust the shown 8D instance in 10 000 calls. All EA trials result in broad distributions of final scores as shown in figures 3.11 to 3.15.

The mathematical description of the charged marble problem

The goal is to place N particles on a circular track and to minimise the total potential energy in a way that equilibrates a repulsive potential between next neighbours with their potential energy on the hilly track (as in fig. V.2). Thus the objective function to be minimised is a weighted sum of two goals

$$F_{\text{obj}}(\vec{x}) = E_{\text{tot}} = \alpha \cdot E_C + \beta \cdot E_G \tag{V.1}$$

where $\vec{x}$ is the list of angles $x_i \in [0, 2\pi]$. E_C is the sum of the repulsive Coulomb potentials between next neighbours:

$$E_C = (x_1 - x_N)^{-1} + \sum_{i=1}^{N-1} (x_{i+1} - x_i)^{-1}. \tag{V.2}$$

The restriction to only next neighbours is a sacrifice owing to a lower computational impact. E_G is the track potential consisting of the potential energy of each "marble on the hilly track":

$$E_G = \sum_{i=1}^{N} m_i h(x_i) \tag{V.3}$$

$$\text{with} \quad h(x) = \sum_{i=1}^{3} A_i(\sin(x + \delta_i) + 1), \tag{V.4}$$

$$\text{whereby} \quad A_1 = 0.6; \quad A_2 = 1; \quad A_3 = 0.8;$$

$$\delta_1 = 0; \quad \delta_2 = \frac{40 \cdot 2\pi}{360}; \quad \delta_3 = \frac{10 \cdot 2\pi}{360};$$

$$m_i = 1 - 0.8 \cdot \frac{(i-1)}{(N-1)}; \quad \text{with} \quad i \in [1, 2, ..., N]$$

This $h(x)$ creates the altitude profile forming the black and white track background shown in figure V.2. The masses m_i are equidistant and cover the interval $[0.2, 1]$. Setting the weighting ratio to $\alpha = 1$ and $\beta = 2.5$ brings the two goals into a competition so that the global optimum is neither a state of all marbles being collected in the deepest valley nor one where the hill track just introduces a tiny irregularity into an almost even angle distribution.

The relevance of the charged marble problem in the context of resonator optimisation

The decision on which EA to apply to the resonator optimisation task was based to a substantial part on the algorithm performance on the charged marble problem because of the following similarities concerning the characteristics of the two objective functions, the energy E_{tot} in the marble problem, and the peak sound pressure p_{max} in the resonator tuning:

- The diverging energy barriers when one charged particle crosses another are equivalent to the deep canyons in p_{max} when interacting resonances quench each other.
- The smooth and sinusoidal energy ondulations created when marbles are pushed over hills and through valleys are responsible for a similarly smooth and hilly background structure of the objective function as when the amplitude of a particular resonance grows and shrinks as a consequence of varying a resonator's design parameters.
- The objective functions in both cases are smooth and nonseparable.
- A combinatorial aspect exists in both cases: marbles and valleys have to be combined in the test problem, the right stiffnesses, masses, and wavelengths have to be combined in the resonator.

Exploiting these similarities for making the right EA choice means leveraging the available knowledge about the optimisation problem for fighting the no free lunch theorem [512, 519].

Comparison to other visualisable test problems

In order to explain why it is thought that this type of test problem fills a rather empty gap, a comparison with several other visualisable problems was compiled for a conference publication [428]. As it is of secondary importance in the current context, the main conclusions are only given very briefly here and only in connection with the three most illustrative comparisons.

The first example is the Rastrigin function which has been discussed with a visualisation scheme above. The plots do not only reveal the solution quality. Another feature is that it is evident for the observer how the chromosome should be modified in order to reach the global minimum. The distance of the current chromosome from the global minimum can be inferred, and even more importantly, because one knows how many dots have to be pushed over how many hills, one knows about the topology and the difficulty of the way lying still ahead in the search space. But this is all only possible in the unrotated separable case. In the nonseparable case of the rotated Rastrigin function (as in [444]), equivalent plots can of course still be drawn, but most of the meaning is lost. The grey 1D potential at the bottom would have to be erased, marking also the impossibility to make any further conclusions about what lies ahead in the search space. While in the unrotated case one knows whether

the incremental move of one dot will cost or gain energy, it is impossible in the unrotated case. Expressing it still in another way, comparing two pictures before and after a mutation, one cannot tell any more from the plot whether the mutation was beneficial or not. The observer loses the position of being an all-knowing observer who can judge whether an optimiser does the right things or not.

The second example is the cantilever beam optimisation presented in the context of the application of evolution strategies to bionics on the website of Rechenberg's group [378]. It is illustrated in figure V.3. The comparison of high- with low-quality solutions in the figure reminds us that beauty and efficiency is often correlated in problems of light-weight architecture and structural stability faced and solved by nature and humans. It is no coincidence but a result of evolution, that we have a well-trained eye to distinguish stable from weak branches in trees and need just fractions of a second for such decisions. This fact is the root why this truss construction optimisation problem is easy to solve for us, e. g. by iteratively identifying the node in the oddest position and pushing it into a better place. This again creates the perfect situation – from the point of view of the EA developer needing a fast-feedback test problem – that the human observer is omniscient and can instantly judge the gains and mistakes made by an optimisation algorithm working on the task. That this applies to a multimodal problem with no less than twelve degrees of freedom, sounds remarkable at first. The only problem is that the problem is not at all challenging for standard EAs. As the flexible six nodes are "pulled" together to form an evenly shaped curve, the optima can be reached by purely downhill runs starting with any low-quality solution. Are the seven local minima forming a regular pattern? Is there a common valley with only small and smooth ondulations at the bottom for their separation, a 1D chain of basins like a staircase? Almost. Going from the global optimum through the six other local minima, always all nodes have to be pushed upwards a little bit. Since each time one of them has to do a larger jump, it can be said that the chain of basins takes $\approx 90°$ angle turns spiralling through the search space. As the truss construction has a narrow neck when the upwards flipping node has to cross the straight line of upper nodes, there are high barriers of relatively heavy-weight solutions separating the basins. But for sufficiently large mutation steps this is much less of an obstacle than for deterministic gradient-followers. So the low difficulty level comes from a low number of basins forming a simple edgy chain in a landscape which elsewhere has only smooth and monotonically mounting slopes.

The third example is the FM-wave synthesis problem from the CEC-2011 collection of real-world problems [102]. FM is short for frequency modulation and it means that a periodic signal's frequency is modulated by the output of another oscillating signal. Given a signal

$$s(t) = A_1 \sin\Big(\omega_1 t + A_2 \sin\left(\omega_2 t + A_3 \sin(\omega_3 t)\right)\Big)$$

stretching over an interval $t \in [0, T]$, a function fitting problem can be defined: find the parameter set $\{A_1, \omega_1, A_2, \omega_2, A_3, \omega_3\}$ that is able to reproduce the original signal, i.e. that minimises the area between the original and the trial signal to zero. Alternatively, on a discretised grid of check points t_k with $k \in [1, ..., K]$ the candidate solution can be required to match the original signal in a least-squares sense. This

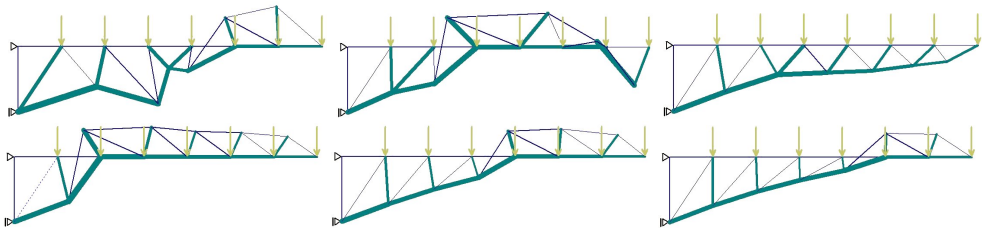


Figure V.3 Cantilever optimisation by ES
These truss cantilever images are snapshots taken from one of the evolution strategy demo programs hosted on the website of the bionics group of the Technical University Berlin [378]. In this parametrised 2D truss problem, the weight of a cantilever has to be minimised. The fixed support points are indicated by triangles and the fixed external weight loads by the yellow-beige arrows. The forces on the truss elements are calculated, and then a weight of each truss is calculated based on what cross section the corresponding truss (green) or rope (dark blue) must have in order not to fail. What has to be tuned are the x- and y-positions of six of the seven nodes in the lower row. (Obviously the lower nodes are allowed to flip above the upper row.) The pictures represent various stages of different optimisation runs using a particular setup of branching conventional evolution strategies. The upper row of snapshots shows two low-quality solutions and a near-optimal one. The lower row shows locally optimal solutions. The optimised light-weight constructions are perceived as aesthetic and even-shaped. An even S-line of trusses has to be formed in the lower row, and an even line of ropes has to form in the flipped versions. Disturbing the even curves by shifting one single node up-, down-, or sidewards, raises the weight of the construction. As with rubber strings, each node is pulled (by the objective function "weight") towards the line formed by its neighbours. Therefore, the difficulty level of this optimisation task is quite low. The number of seven local minima is low considering the twelve degrees of freedom.

is the setting of the FM-synthesis problem formulated in the CEC-2011 collection [102]. A perfect visualisation mode is possible for this problem, as shown in figure V.4.

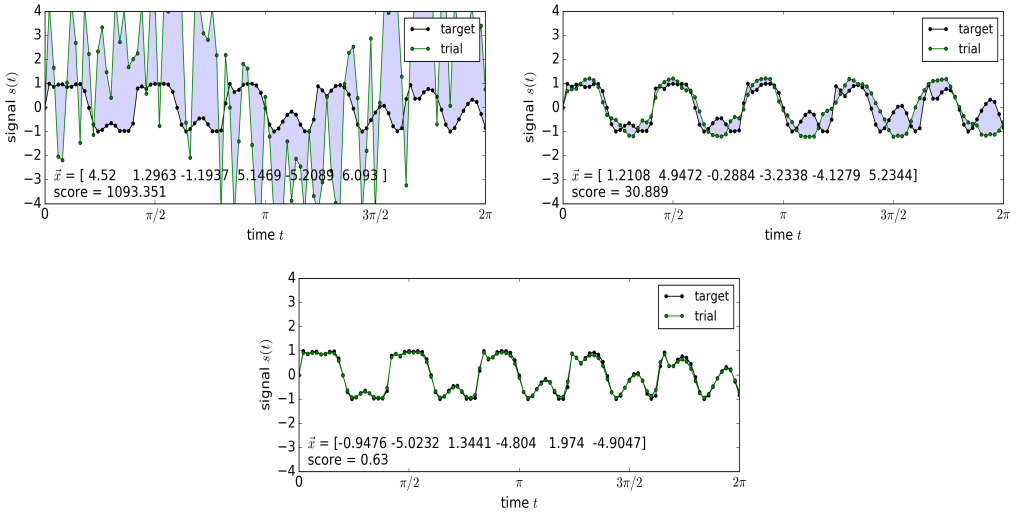


Figure V.4 The FM wave synthesis problem
The FM wave synthesis problem (problem no. 1 of the CEC-2011 list of real-world benchmarking problems [102]) shows a sufficient degree of difficulty to become useful for evaluating stochastic optimisers. The goal is to fit a target signal from 3 nested sine functions by tuning the set of amplitudes and frequencies. With the help of plots like depicted above one can intuitively estimate relative fitness differences within fractions of a second. However, the relationship between the setting of the tuning parameters and the amount of shaded surface cannot be grasped that easily. One does not know in which directions to increment the amplitudes and frequencies (except the first amplitude), and one does not have a feeling about how many and how high potential barriers still have to be surmounted to get from a local to the global optimum.

The problem is challenging. Everybody who has ever used an analogue synthesiser for creating FM sounds and has tried to restore the setting (after all the dials had been randomised) of an interesting sound heard before knows that. Furthermore, the search task has two special features. One is that the different parameters have a different impact. A_1 just scales the function values. All the amplitudes have a lower impact than the frequencies. Of the frequencies ω_3 has the strongest and least predictable impact because of the propagation through all the three sine functions. The other one is that search spaces of different topologies can be created. Depending on the choices made for the target parameters on the one hand, and the window on the t-axis on the other, the search landscape can be tuned from simple and smooth to something almost similar to random noise when high frequencies and a large capture window are used.

The only shortcoming of the problem class is that the human observer is not omniscient in the sense described above and does almost never know in which direction to modify the solution candidate $\vec{x}$ in order to get closer to the global optimum, nor can she or he have an intuitive understanding of what obstacles still lie ahead. This means that the FM-synthesis test problem is perfect for scrolling through several optimisation histories and getting a feeling of the overall optimisation speed and success, but it is not suitable to go down into the detail level of single mutation steps, no information can be gained on that level for judging the usefulness of an algorithm's search moves.

It is hoped that this selection of examples illustrates the conflict apparently arising between visualisability (in the described intuitive sense) and difficulty level. As this conflict is unresolved by the test problems found in literature and discussed in the dedicated conference article [428], the *charged marble test problem* is thought to be a valuable addition to the tool set available for the developer of EAs.

Disadvantage or feature? – structure in the search space

Is the search space spanned by the charged marble problem completely irregular or are there patterns? The situation is as follows: local energy minima are separated by two sorts of barriers. There are low barriers when one marble has to roll over a hilltop and infinitely high barriers (acting like separating walls or fences) arise when marbles have to cross each other (only with a large enough step size these barriers con be surmounted easily). Let's start at the global optimum and let the whole marble pattern rotate in lockstep over the ring track. In the search space this is a straight trajectory parallel to the vector $\vec{1} = (1, 1, ..., 1)$. Only barriers of the low hilly type are seen along that path. Any motion perpendicular to this path will involve marbles crossing each other. This means the infinitely high walls are separating tube-like volumes forming a sort of honeycomb structure. The hill structure inside each tube repeats itself when the marble pattern has rotated by 2π; one has to restrict the search to $x_i \in [0, 2\pi]$ for avoiding that type of repetitiveness.

So it has to be admitted that the charged marble problem is not completely neutral in terms of the spatial distribution of local minima. There is a clear anisotropicity arising from the honeycomb structure of the infinitely high energy barriers. Whether this has to be judged as irrelevant or a severe drawback for a testing or

benchmarking function ultimately hinges on the purpose which the EA application is aiming at. In the exemplary application case of SF resonator optimisation, where masses and stiffnesses are tightly correlated if resonances of substructures have to be matched, it is deemed that there are similar elements of anisotropicity.

Summing up advantages and disadvantages

In short, the benefits of the charged marble problem can be summed up as follows:

- It is a nonseparable multimodal minimisation problem with no apparent and simple regularity in the distribution of maxima and minima.
- It is a continuous optimisation problem with a combinatorial aspect.
- The features defining the difficulty level can be switched on and off or faded in: the amplitude of the track potential, the weight differences of the marbles, the number of marbles.
- The ruggedness of the hills is directly reflected in the objective function and different surface properties can be tried.
- Due to the sum of two types of potential energies the problem can be seen as single- or biobjective (and the relation between the separate objectives via the energy scale is physically meaningful).
- There is a visualisation allowing intuitive perception by human observers, who can judge relative quality differences at a glance.
- The visualisation is at the same time a visualisation of the chromosome $\vec{x}$ as well as the score, it shows geno- and phenotype.
- The observer is (almost) all-knowing and aware of how to modify the genotype in order to improve the quality of the phenotype.
- The above is true even when the solution candidate is still far from the global minimum and energy barriers are lying in between.
- The observer has an intuitive estimation of how difficult the remaining way towards a near-optimal solution still is, i. e. whether there are many or few energy barriers to be surmounted.
- Whole optimisation histories can be plotted like in figure V.5.

On the other hand, there are these shortcomings:

- The intuitive understandability of the visualisation is not possible for too many dimensions.
- It is also lost if the track potential is given a too complex shape.
- It is lost as well if the inter-particle potential is changed.

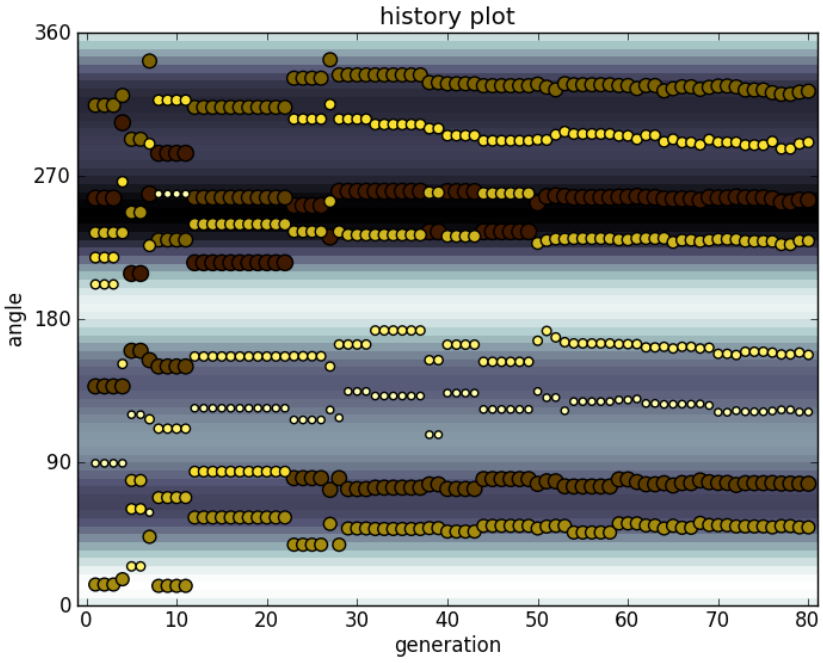


Figure V.5 Optimisation history view of the charged marble problem
This view offers additional insights on how a tested EA conducts the search. Time periods during which only local improvement steps were made are clearly distinguishable. Other periods where several back and forth switches occur between two or more competing solutions are proof that the EA is good at conserving gene pool diversity, i. e. that it is able to locally investigate several hot spots of the population cloud in parallel.

- The intuitive judgement is not quite exact. In cases of pair comparisons where the ordering is different but the energy similar it fails.
- For a new problem instance the global minimum is generally unknown.
- There is some regular patterning of the search space, the objective function's structure is not isotropic. (If the test problem has a type of structure not exhibited by the real-world problem, then this can lead to wrong EA choices.)
- If the weight balance between the two types of energies is modified, one needs to re-learn what good solutions look like by watching many optimisation runs. At other weight ratios it may be much harder or impossible to gain the ability of intuitive judgement at all.

Expanding the problem class

The basic concept behind this problem class opens a wider range of possibilities than just modifying the parameters in the above problem statement and the amount of marbles on the track. The complexity of the problem instances can be scaled, and the characteristics can be tuned by modifications such as:

- increasing the complexity of the ridge line $h(x)$ by introducing finer ripples, edges, cliffs, plateaus, etc.,

- a straight track of finite length instead of a circle,
- context of multi-objective optimisation: possibility of a third goal, e.g. by addition of another gravitation pulling towards the bottom of the picture, or addition of an off-track repelling particle at a fixed position
- noise: individual marble masses or charges vary from generation to generation,
- context of EAs for dynamic or online optimisation: dynamic from the side of the track morphing into another shape over time,
- similarly, dynamic can also come from the side of the inter-particle potential by time-variation of its amplitude or shape,
- two ring tracks with different characteristics as shown in figure V.6 (left),
- marbles not fixated to circular tracks but moving within a rectangular plane with hilly background (as in fig. V.6).
- Finding the optimal highest-density packing configuration of apples in a box is relatively easy, but with twisted wooden branches it is much more difficult. Therefore, building on the idea of the rectangular plane with hilly background, the optimisation task's complexity can still be substantially raised if molecules with less symmetry are to be placed ideally instead of single particles. The first option is pairs of particles (connecting the marbles with little rods), and L-shaped triplets could make it really difficult (see fig. V.6).

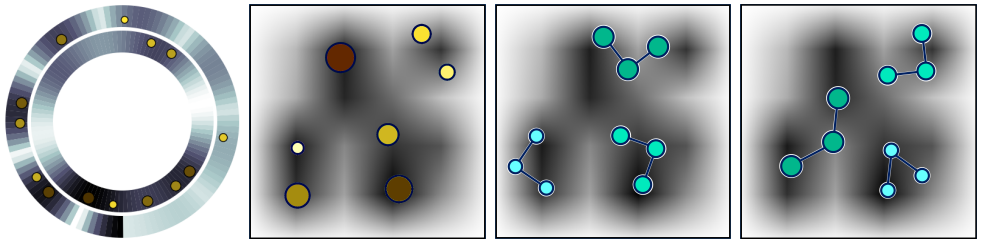


Figure V.6 Changing problem characteristics and complexity
Illustration of some ideas for changing problem characteristics and complexity: edges and cliffs in the track potential, two neighbouring tracks, 2D-plane instead of 1D rail, molecules of reduced symmetry instead of particles.

Lists of symbols and abbreviations

List of Latin quantity symbols

Symbol	Description
a, b	generic parameters
A_i	generic parameters, used for wave amplitudes
E	energy
f_{obj}	objective function
h	height, elevation

K	generic integer parameter
m	mass
$\mathbf{M}$	rotation matrix
N	number of particles
n	search space dimension
P	probability
r	radius
$s(t)$	signal
t	time
x_i	input vector component
$\vec{x}$	input vector, point in search space, solution candidate
x, y	generic scalar function input variables
$\vec{z}$	intermediate point, i. e. output of transformation and input to test function kernel (CEC-2005 notation)

List of Greek quantity symbols

Symbol	**Description**
α, β	generic parameters
Δ	additive offset to f_{obj} (CEC-2005 notation)
δ_i	phase offsets
θ	time step
Ω	translation vector and effective location of global optimum (CEC-2005 notation)
ω	angular frequency

List of abbreviations

Abbreviation	**Description**
CEC	IEEE Congress on Evolutionary Computation
CMA-ES	evolution strategy with covariance matrix adaptation
DE	differential evolution
EA	evolutionary algorithm
ES	evolution strategy
FM	frequency modulation
GA	genetic algorithm
PSO	particle swarm optimisation

Appendix W

Listing of optimised parameter sets

This appendix contains listings of the tuned design parameters for geometries A, B, and C as well as the West-Howlett geometry. In the case of geometry A where many cases are described in chapter 5 only two selected setups are explicitly listed here. The parameter listings are complete in the sense that together with the parametrisation schematics displayed in chapter 5 they fully determine the resonator geometries. Parameters having been subject to optimisation are designated by shaded fields.

Table W.1 Optimised parameter sets for the West-Howlett geometry

case	manually tuned	EA-tuned
id	59.2	59.2
ih	78.0	80.898
pit	3.0	3.0
pih	25.0	25.0
pipos	-0.86	0.633
ept	0.5	0.5
ext_uh	40.0	45.583
ext_lh	20.0	37.213
wt	2.4	1.501
sh1	3.0	3.0
silpos	0.5	0.5
sh2	1.0	1.0
sh3	3.0	3.0
sg	0.5	0.5
gap	7.0	4.833
pt	15.0	14.088
pwt1	2.0	3.041
pwt2	2.0	2.126
tr	3.55	3.55
twt	1.05	1.05
icr	6.5	6.5
tl	100.0	100.0

Table W.2 Optimised parameter sets for geometry A

case	140	212
id	61.241	60.0
ih	60.118	75.34
pit	3.0	3.0
pih	25.0	25.0
pipos	-0.663	-0.694
ept	0.5	0.5
ext_uh	41.173	57.904
ext_lh	32.715	30.089
wt	2.005	2.4
gap_uh	8.853	7.715
gap_lh	3.632	2.082
pt_uh	14.012	14.651
pt_lh	10.478	21.943
pwt1_lh	3.575	2.401
pwt1_lh	3.575	2.401
pwt2_uh	5.047	6.989
pwt2_lh	2.281	7.572
tr	5.0	5.0
twt	2.0	2.0
bpt	6.0	6.0
sh1	3.126	7.777
sh2	10.0	10.0
silext	3.194	8.057

Table W.3 Optimised parameter sets for geometry B

case	33	34	35
id	59.2	59.2	59.2
ih	40.0	166.003	40.0
pit	3.0	3.0	3.0
pih	25.0	25.0	25.0
ept	0.5	0.5	0.5
ext	16.864	24.603	16.87
wt	2.4	2.4	2.4
sh	4.565	2.832	4.88
gap	11.044	3.01	10.202
pwt1	1.797	16.0	2.146
pwt2	1.0	3.0	1.0
bpt	11.543	15.048	11.569
fxw	5.0	5.0	5.0

Table W.4 Optimised parameter sets for geometry C

case	12	13	85	87	92	96
bpt_uh	–	–	37.621	32.138	–	–
bpt_lh	47.284	25.678	16.455	26.418	34.925	43.449
bpw_uh	–	–	49.218	31.465	–	–
bpw_lh	36.896	48.046	38.247	34.025	39.915	34.455
ept	0.5	0.5	0.5	0.5	0.5	0.5
ext_uh	–	–	46.89	45.993	–	–
ext_lh	12.347	22.772	34.423	47.761	27.016	36.189
fir1_uh	–	–	1.96	6.594	–	–
fir1_lh	1.731	2.196	2.967	1.101	3.382	6.117
fir2_uh	–	–	1.23	1.467	–	–
fir2_lh	1.831	0.972	2.685	2.632	2.311	1.218
fir3_uh	–	–	1.2	1.2	–	–
fir3_lh	1.2	1.2	1.2	1.2	1.2	1.2
fir4	1.0	1.0	1.0	1.0	1.0	1.0
gap_uh	–	–	10.125	8.475	–	–
gap_lh	5.045	5.023	5.418	12.984	6.859	5.125
ih	60.293	62.235	75.472	67.953	86.313	99.129
ir	29.6	29.6	29.6	29.6	29.6	29.6
iros_uh	–	–	3.507	2.486	–	–
iros_lh	3.192	0.862	4.651	1.777	3.146	3.353
pid	3.0	3.0	3.0	3.0	3.0	3.0
pih	25.0	25.0	25.0	25.0	25.0	25.0
pipos	-0.7	-0.869	-0.649	-0.915	-0.882	-0.716
pwt1_uh	–	–	5.299	3.347	–	–
pwt1_lh	4.719	4.594	1.734	5.312	4.797	2.932
pwt2_uh	–	–	2.245	0.525	–	–
pwt2_lh	3.695	5.777	0.829	5.579	3.98	1.612
pwt3_uh	–	–	2.696	2.184	–	–
pwt3_lh	3.164	1.781	3.935	6.932	2.58	0.728
pwt4_uh	–	–	3.206	2.199	–	–
pwt4_lh	0.521	4.319	1.082	4.358	2.518	0.901
pwt5_uh	–	–	4.03	1.83	–	–
pwt5_lh	1.862	2.125	3.797	1.034	1.913	2.069
pwt6_uh	–	–	3.672	3.067	–	–
pwt6_lh	3.834	2.331	0.942	4.382	2.023	4.634
pwpA_uh	–	–	0.648	0.367	–	–
pwpA_lh	0.309	0.522	0.46	0.207	0.255	0.159
pwpAA_uh	–	–	0.206	0.162	–	–
pwpAA_lh	0.85	0.125	0.809	0.807	0.629	0.22
pwpAAA_uh	–	–	0.217	0.827	–	–
pwpAAA_lh	0.708	0.191	0.5	0.681	0.209	0.26
pwpAB_uh	–	–	0.392	0.355	–	–
pwpAB_lh	0.461	0.659	0.638	0.164	0.685	0.711
pwpB_uh	–	–	0.163	0.666	–	–
pwpB_lh	0.105	0.567	0.597	0.615	0.252	0.415
pwpBA_uh	–	–	0.88	0.227	–	–
pwpBA_lh	0.714	0.557	0.752	0.261	0.387	0.31
sh1_uh	–	–	2.344	3.66	–	–
sh1_lh	1.46	1.508	0.525	2.901	1.472	4.161
sh2_uh	–	–	0.788	4.429	–	–
sh2_lh	3.946	1.8	3.423	1.656	3.745	3.883
swi_uh	–	–	3.729	3.85	–	–
swi_lh	0.537	1.347	0.988	3.567	0.942	0.749
swo_uh	–	–	4.788	1.651	–	–
swo_lh	2.943	4.405	4.445	4.478	3.739	2.346
wt	2.4	2.4	2.4	2.4	2.4	2.4

Bibliography

[1] *ACT / ESA global optimisation solvers*. `http://www.esa.int/gsp/ACT/inf/op/globopt/solvers.htm`. (visited on 2013/10/17).

[2] M. Adachi, Y. Akishige, T. Asahi, K. Deguchi, K. Gesi, K. Hasebe, T. Hikita, T. Ikeda, Y. Iwata, M. Komukae, T. Mitsui, E. Nakamura, N. Nakatani, M. Okuyama, T. Osaka, A. Sakai, E. Sawaguchi, Y. Shiozaki, T. Takenaka, K. Toyoda, T. Tsukamoto, and T. Yagi. *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*. Ed. by Y. Shiozaki, E. Nakamura, and T. Mitsui. NS III/36A1. Springer, 2002.

[3] Kodjo Agbossou and Jean-Luc Dion. "Reacteur a cavitation acoustique". Patent EP/1087837 B1. Nov. 2003.

[4] I. Akhatov, U. Parlitz, and W. Lauterborn. "Towards a theory of self-organization phenomena in bubble-liquid mixtures". In: *Physical Review E* 54.5 (Nov. 1996), pp. 4990–5003.

[5] Carsten Lindegaard Andersen. *personal corresponding*. Ortofon A/S, Nakskov, Denmark, 2009.

[6] Mark Anderson. *Bubble Fusion Bubbles Up Again - IEEE Spectrum*. IEEE Spectrum. (visited on 2014/10/07). July 2013.

[7] *ANSYS Academic research, Release 14.0, Help System*. ANSYS, Inc. Southpointe, 275 Technology Drive, Canonsburg, PA 15317, USA.

[8] *ANSYS Academic research, Release 14.0, Help System, Theory Reference*. ANSYS, Inc. Southpointe, 275 Technology Drive, Canonsburg, PA 15317, USA.

[9] *ANSYS Academic research, Release 14.0, Help System, Verification Manual*. ANSYS, Inc. Southpointe, 275 Technology Drive, Canonsburg, PA 15317, USA.

[10] R. E. Apfel, S. C. Roy, and Y.-C. Lo. "Prediction of the minimum neutron energy to nucleate vapor bubbles in superheated liquids". In: *Physical Review A* 31.5 (May 1985), pp. 3194–3198.

[11] Vijay H. Arakeri. "Sonoluminescence and bubble fusion". In: *Current Science* 85.7 (Oct. 2003), pp. 911–916.

[12] Brian C Archambault. "Ascertaining directional information from incident nuclear radiation in the acoustically tensioned metastable fluid detector system". master thesis. West Lafayette, Indiana: Purdue University, May 2010.

[13] Antonio Arnau. *Piezoelectric Transducers and Applications*. Springer Science & Business Media, Oct. 2008.

[14] Peter Atkins and Julio de Paula. *Atkins' Physical chemistry*. Oxford; New York: Oxford University Press, 2006.

[15] W. Atmar. "On the rules and nature of simulated evolutionary programming". In: *Proc. First Ann. Conf. on Evolutionary Programming*. LaJolla: Evolutionary Programming Society, 1992, pp. 17–26.

[16] G. Audi, A. H. Wapstra, and C. Thibault. "The Ame2003 atomic mass evaluation: (II). Tables, graphs and references". In: *Nuclear Physics A*. The 2003 NUBASE and Atomic Mass Evaluations 729.1 (Dec. 2003), pp. 337–676.

[17] A. Auger and N. Hansen. "A restart CMA evolution strategy with increasing population size". In: *The 2005 IEEE Congress on Evolutionary Computation, 2005*. Vol. 2. Sept. 2005, 1769–1776 Vol. 2.

[18] Anne Auger and Nikolaus Hansen. *Tutorial – CMA-ES (Covariance Matrix Adaptation Evolution Strategies)*. `http://www.lri.fr/%7Ehansen/gecco2011-CMA-ES-tutorial.pdf`. (visited on 2017/11/23). Dublin, Ireland, July 2011.

[19] Anne Auger and Olivier Teytaud. "Continuous Lunches Are Free Plus the Design of Optimal Optimization Algorithms". In: *Algorithmica* 57.1 (May 2010), pp. 121–146.

[20] T. Bäck, U. Hammel, and H.-P. Schwefel. "Evolutionary computation: comments on the history and current state". In: *IEEE Transactions on Evolutionary Computation* 1.1 (Apr. 1997), pp. 3–17.

[21] Thomas Bäck, Frank Hoffmeister, and Hans-Paul Schwefel. "A Survey of Evolution Strategies". In: *Proceedings of the Fourth International Conference on Genetic Algorithms*. Morgan Kaufmann, 1991, pp. 2–9.

[22] Kenneth B. Bader, Jason L. Raymond, Joel Mobley, Charles C. Church, and D. Felipe Gaitan. "The effect of static pressure on the inertial cavitation threshold". In: *The Journal of the Acoustical Society of America* 132.2 (Aug. 2012), pp. 728–737.

[23] A. Ballato. "Modeling piezoelectric and piezomagnetic devices and structures via equivalent networks". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 48.5 (Sept. 2001), pp. 1189–1240.

[24] Arthur Ballato. *Equivalent circuits for resonators and transducers driven piezoelectrically*. Research and Development Technical Report SLCET-TR-90-12. Oct. 1990.

[25] M. Barbaglia, P. Florido, R. Mayer, and F. Bonetto. "Search of Fusion Reactions During the Cavitation of a Single Bubble in Deuterated Liquids". In: *Physica Scripta* 72.1 (Jan. 2005), p. 75.

[26] Bradley P. Barber, Robert A. Hiller, Ritva Löfstedt, Seth J. Putterman, and Keith R. Weninger. "Defining the unknowns of sonoluminescence". In: *Physics Reports* 281.2 (Mar. 1997), pp. 65–143.

[27] Bradley P. Barber and Seth J. Putterman. “Observation of synchronous picosecond sonoluminescence”. In: *Nature* 352.6333 (July 1991), pp. 318–320.

[28] Bradley P. Barber, C. C. Wu, Ritva Löfstedt, Paul H. Roberts, and Seth J. Putterman. “Sensitivity of sonoluminescence to experimental parameters”. In: *Physical Review Letters* 72.9 (Feb. 1994), pp. 1380–1383.

[29] Alexander Bass, Seth Putterman, Barry Merriman, and Steven J. Ruuth. “Symmetry reduction for molecular dynamics simulation of an imploding gas bubble”. In: *Journal of Computational Physics* 227.3 (Jan. 2008), pp. 2118–2129.

[30] Alexander Bass, Steven J. Ruuth, Carlos Camara, Barry Merriman, and Seth Putterman. “Molecular Dynamics of Extreme Mass Segregation in a Rapidly Collapsing Bubble”. In: *Physical Review Letters* 101.23 (Dec. 2008), p. 234301.

[31] Klaus-Jürgen Bathe. *Finite-Elemente-Methoden*. Berlin, Heidelberg, New York, Tokyo: Springer, 1986.

[32] F. D. Becchetti. “Evidence for Nuclear Reactions in Imploding Bubbles”. In: *Science* 295.5561 (Mar. 2002), pp. 1850–1850.

[33] C. R. Bell, N. P. Oberle, W. Rohsenow, N. Todreas, and C. Tso. “Radiation-Induced Boiling in Superheated Water and Organic Liquids”. In: *Nucl. Sci. Eng., v. 53, no. 4, pp. 458-465* (Apr. 1974).

[34] Paul M. Bellan. *Spheromaks*. `http://ve4xm.caltech.edu/Bellan_plasma_page/spheroma.html`. (visited on 2015/03/25). Pasadena, CA.

[35] C. A. Bellera, M. Julien, and J. Hanley. “Normal Approximations to the Distributions of the Wilcoxon Statistics: Accurate to What N? Graphical Insights”. In: *Journal of Statistics Education* 18.2 (2010).

[36] Abraham I. Beltzer. *Variational and finite element methods: A symbolic computation approach*. New York Berlin Heidelberg: Springer-Verlag, 1990.

[37] E. Benton, N. Lindy, W. H. Beasley, and D. Petersen. “Lightning and Extensive Air Showers of Cosmic Rays”. In: *AGU Fall Meeting Proceedings*. San Francisco, CA, USA, Dec. 2011.

[38] Steffen Bergweiler. “Körperoszillation und Schallabstrahlung akustischer Wellenleiter unter Berücksichtigung von Wandungseinflüssen und Kopplungseffekten: verändern Metalllegierung und Wandungsprofil des Rohrresonators den Klang der labialen Orgelpfeife?” doctoral thesis. Potsdam: Universität Potsdam, 2006.

[39] C. W. Bert. “Material damping: An introductory review of mathematic measures and experimental technique”. In: *Journal of Sound and Vibration* 29.2 (July 1973), pp. 129–153.

[40] Hans-Georg Beyer. “The simple genetic algorithm – foundations and theory [Book Reviews]”. In: *IEEE Transactions on Evolutionary Computation* 4.2 (July 2000), pp. 191–192.

[41] Hans-Georg Beyer. “Toward a theory of evolution strategies: Some asymptotical results from the $(1, +\lambda)$-theory”. In: *Evol. Comput.* 1.2 (June 1993), pp. 165–188.

[42] Hans-Georg Beyer and Hans-Paul Schwefel. “Evolution strategies – A comprehensive introduction”. en. In: *Natural Computing* 1.1 (Mar. 2002), pp. 3–52.

[43] Hans-Georg Beyer, Hans-Paul Schwefel, and Ingo Wegener. *How to analyse evolutionary algorithms*. Tech. rep. CI-139/02. Dortmund: Department of Computer Science/XI, University of Dortmund, Aug. 2002.

[44] *Bicycle performance*. `https://en.wikipedia.org/w/index.php?title=Bicycle_performance&oldid=776654403`. Page Version ID: 776654403 (visited on 2017/04/22). Apr. 2017.

[45] Mauro Birattari, Zhi Yuan, Prasanna Balaprakash, and Thomas Stützle. “F-Race and Iterated F-Race: An Overview”. In: *Experimental Methods for the Analysis of Optimization Algorithms*. Ed. by Thomas Bartz-Beielstein, Marco Chiarandini, Luís Paquete, and Mike Preuss. Springer Berlin Heidelberg, Jan. 2010, pp. 311–336.

[46] R. E. D. Bishop and D. C. Johnson. *The Mechanics of Vibration*. Cambridge University Press, 1979.

[47] V. Bityurin, A. Bykov, V. Velikodny, A. Dyrenkov, and B. Tolkunov. “Theoretical and experemental reseach of shock wave infuence on deuterium porous liquid (Theoretical and experimental investigation of the effect of shock waves on porous deuterated liquid)”. In: *Fiziko-Khimicheskaja Kinetika v Gazovoy Dinamike* 6 (2008).

[48] M.S. Blanter, I.S. Golovin, H. Neuhäuser, and H.-R. Sinnig. *Internal Friction in Metallic Materials - A Handbook*. Vol. 90. Springer Series in Materials Science. Springer, 2007.

[49] Yu. I. Bobrovnitskii. “Hysteretic damping and causality”. In: *Acoustical Physics* 59.3 (May 2013), pp. 253–256.

[50] Terry B. Bollinger. “Ultra Cavitation – An Outline of Theoretical and Experimental Issues”. 2416 Branch Oaks Lane, Flower Mound, Texas 75028, Jan. 1993.

[51] A. Bolufe-Rohler, S. Estevez-Velarde, A. Piad-Morffis, S. Chen, and J. Montgomery. “Differential evolution with thresheld convergence”. In: *2013 IEEE Congress on Evolutionary Computation (CEC)*. June 2013, pp. 40–47.

[52] Hans-Stephan Bosch and Wendelstein 7-X Team. “Wendelstein 7-X – a Technology Step towards Demo”. In: *Plasma and Fusion Research* 5 (2010), S1002–1 – S1002–7.

[53] A. Bougaev, J. Walters, T. Jevremovic, M. Bertodano, F. Clikeman, E. Merrit, S. Revankar, and L. H. Tsoukalas. “Tritium Evidence in Acoustic Cavitation Nuclear Emission Experiments (Draft)”. `http://newenergytimes.com/v2/bubblegate/2004/2004-TsoukalasL-TritiumEvidence-Draft.pdf`. (visited on 2016/04/04). West Lafayette, Indiana, 2004.

[54] Ilhem Boussaïd, Julien Lepagnot, and Patrick Siarry. "A survey on optimization metaheuristics". In: *Information Sciences* 237 (July 2013), pp. 82–117.

[55] H. J. Bremermann. "Numerical optimization procedures derived from biological evolution processes". In: *Cybernetic Problems in Bionics*. New York: Gordon and Breach, pp. 597–615.

[56] Christopher Earls Brennen. *Cavitation and Bubble Dynamics*. Jan. 1995.

[57] Michael P. Brenner, Sascha Hilgenfeldt, and Detlef Lohse. "Single-bubble sonoluminescence". In: *Reviews of Modern Physics* 74.2 (May 2002), pp. 425–484.

[58] J. Brest, S. Greiner, B. Boskovic, M. Mernik, and V. Zumer. "Self-Adapting Control Parameters in Differential Evolution: A Comparative Study on Numerical Benchmark Problems". In: *IEEE Transactions on Evolutionary Computation* 10.6 (Dec. 2006), pp. 646–657.

[59] Janez Brest. "Constrained Real-Parameter Optimization with ε-Self-Adaptive Differential Evolution". In: *Constraint-Handling in Evolutionary Optimization*. Ed. by Efrén Mezura-Montes. Vol. 198. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 73–93.

[60] Malcolm W. Browne. "New Shot at Cold Fusion By Pumping Sound Waves Into Tiny Bubbles". In: *The New York Times* (Dec. 1994).

[61] Jason Brownlee. *Clever Algorithms*. `http://www.cleveralgorithms.com/`. (visited on 2013/10/17).

[62] Jason Brownlee. *Clever Algorithms: Nature-Inspired Programming Recipes*. LuLu, June 2012.

[63] William Bugg. "Affidavit of Dr. William Bugg". `http://newenergytimes.com/v2/bubblegate/Aff/Bugg.pdf`. (visited on 2016/04/24). Jan. 2008.

[64] William Bugg. "Report on Activities on June 6-7 Visit". `http://newenergytimes.com/v2/bubblegate/2006/2006BuggtoTaleyarkhan.pdf`. June 2006.

[65] P. D. S. Burnett, D. M. Chambers, D. Heading, A. Machacek, W. C. Moss, S. Rose, M. Schnittker, R. W. Lee, P. Young, and J. S. Wark. "Modeling a sonoluminescing bubble as a plasma". In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 71.2-6 (Oct. 2001), pp. 215–223.

[66] Adam Butt. "Acoustic inertial confinement fusion: characterization of reaction chamber". master thesis. West Lafayette, Indiana: Purdue University, Dec. 2005.

[67] C. G. Camara, S. D. Hopkins, K. S. Suslick, and S. J. Putterman. "Upper Bound for Neutron Emission from Sonoluminescing Bubbles in Deuterated Acetone". In: *Physical Review Letters* 98.6 (Feb. 2007), p. 064301.

[68] M. Canavan, C. Dugan, and A. Villano. *Nuclear Processes in Bubble Fusion*. NYSERDA report. Troy, NY: Rensselaer Polytechnic Institute, 2003.

[69] Silvina Cancelos. “Effect of acoustically-induced pressures on the permeability of a bullfrog urinary bladder”. doctoral thesis. Troy, NY: Rensselaer Polytechnic Institute, May 2005.

[70] Silvina Cancelos. *personal corresponding*. 2007.

[71] G. Cao, S. Danworaphong, and G. J. Diebold. “A search for laser heating of a sonoluminescing bubble”. In: *The European Physical Journal Special Topics* 153.1 (Jan. 2008), pp. 215–221.

[72] Marcel Caraciolo. *pypso – The Python implementation of Particle Swarm Optimization (PSO) Toolbox*. `https://github.com/marcelcaraciolo/pypso`. (visited on 2014/11/24).

[73] Fabio Cardone, Giovanni Cherubini, and Andrea Petrucci. “Piezonuclear neutrons”. In: *Physics Letters A* 373.8-9 (Feb. 2009), pp. 862–866.

[74] Edwin Cartlidge. “Italian Government Slams Brakes on ’Piezonuclear’ Fission”. In: *Science News* (June 2012).

[75] Marco Caserta and Stefan Voß. “Metaheuristics: Intelligent Problem Solving”. In: *Matheuristics*. Ed. by Vittorio Maniezzo, Thomas Stützle, and Stefan Voß. Annals of Information Systems 10. Springer US, Jan. 2010, pp. 1–38.

[76] Frédéric Caupin, Arnaud Arvengas, Kristina Davitt, Mouna El Mekki Azouzi, Kirill I. Shmulovich, Claire Ramboz, David A. Sessoms, and Abraham D. Stroock. “Exploring water and other liquids at negative pressure”. In: *Journal of Physics: Condensed Matter* 24.28 (2012), p. 284110.

[77] Matteo Ceriotti and Massimiliano Vasile. “MGA trajectory planning with an ACO-inspired algorithm”. In: *Acta Astronautica* 67.9 (Nov. 2010). arXiv: 1104.4668 [cs, math], pp. 1202–1217.

[78] L. Chacón, G. H. Miley, D. C. Barnes, and D. A. Knoll. “Energy gain calculations in Penning fusion systems using a bounce-averaged Fokker-Planck model”. In: *Physics of Plasmas (1994-present)* 7.11 (Nov. 2000), pp. 4547–4560.

[79] Kenneth Chang. “Congress Asks Purdue for Fusion Claim Findings”. In: *The New York Times* (Mar. 2007).

[80] Kenneth Chang. “Experts Say New Desktop Fusion Claims Seem More Credible”. In: *The New York Times* (Mar. 2004).

[81] Kenneth Chang. “Practical Fusion, or Just a Bubble?” In: *The New York Times* (Feb. 2007).

[82] Kenneth Chang. “Researcher Cleared of Misconduct, but Case Is Still Murky”. In: *The New York Times* (Feb. 2007).

[83] Kenneth Chang. “Tiny Bubbles Implode With the Heat of a Star”. In: *The New York Times* (Mar. 2005).

[84] J D Cheeke. “Single-bubble sonoluminescence: "bubble, bubble toil and trouble"”. In: *Canadian Journal of Physics* 75.2 (Feb. 1997), pp. 77–98.

[85] J. David N. Cheeke. *Fundamentals and Applications of Ultrasonic Waves*. CRC Press INC, 2002.

[86] P. Roy Chowdhury and D. N. Basu. "Nuclear matter properties with the re-evaluated coefficients of liquid drop model". In: *Acta Physica Polonica B* 37.6 (June 2006). arXiv: nucl-th/0408013, pp. 1833–1846.

[87] Deborah D. L. Chung. *Composite Materials*. Engineering Materials and Processes. Springer, 2003.

[88] Donald D. Clayton. *Principles of Stellar Evolution and Nucleosynthesis*. University of Chicago Press, 1968.

[89] Hannah Cohen. *Sun Layers*. `http://fusedweb.llnl.gov/CPEP/Chart_Pages/5.Plasmas/Sunlayers.html`. fusedweb.llnl.gov. (visited on 2014/02/07).

[90] Charles J. Colbourn. "The complexity of completing partial Latin squares". In: *Discrete Applied Mathematics* 8.1 (Apr. 1984), pp. 25–30.

[91] A. Corana, M. Marchesi, C. Martini, and S. Ridella. "Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm". In: *ACM Trans. Math. Softw.* 13.3 (Sept. 1987), pp. 262–280.

[92] S. H. Crandall. "The Hysteretic Damping Model in Vibration Theory". In: *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 205.1 (Jan. 1991), pp. 23–28.

[93] David Crouch and Andrew Spurr. "Heat Generating Apparatus". Patent WO/2010/070271 (A1). June 2010.

[94] L. A. Crum. *Sonochemistry and Sonoluminescence*. Springer, Dec. 1998.

[95] Lawrence A. Crum. "Resource Paper: Sonoluminescence". In: *The Journal of the Acoustical Society of America* 138.4 (Oct. 2015), pp. 2181–2205.

[96] Lawrence A. Crum. "Sonoluminescence". In: *Physics Today* 47.9 (Sept. 1994), pp. 22–29.

[97] Lawrence A. Crum. "Sonoluminescence and Acoustic Inertial Confinement Fusion". In: Osaka, Japan, 2003.

[98] Joseph C. Culberson. "On the futility of blind search: An algorithmic view of "no free lunch"". In: *Evol. Comput.* 6.2 (June 1998), pp. 109–127.

[99] Jeppe Seidelin Dam. "The Origin of Sonoluminescence". PhD thesis. Copenhagen: University of Copenhagen, 2006.

[100] Jeppe Seidelin Dam and Mogens T. Levinsen. "Size of the Light-Emitting Region in a Sonoluminescing Bubble". In: *Physical Review Letters* 92.14 (Apr. 2004), p. 144301.

[101] Swagatam Das, Sayan Maity, Bo-Yang Qu, and P.N. Suganthan. "Real-parameter evolutionary multimodal optimization – A survey of the state-of-the-art". In: *Swarm and Evolutionary Computation* 1.2 (June 2011), pp. 71–88.

[102] Swagatam Das and Ponnuthurai Nagaratnam Suganthan. *Problem Definitions and Evaluation Criteria for the CEC 2011 Competition on Testing Evolutionary Algorithms on Real-World Optimization Problems*. Tech. rep. Kolkata, India: Jadavpur University, Dec. 2010.

[103] Richard Dawkins. *The selfish gene*. Oxford University Press, 1976.

[104] Kenneth De Jong. "Evolutionary computation: a unified approach". In: *Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference companion*. GECCO Companion '12. New York, NY, USA: ACM, 2012, pp. 737–750.

[105] Kenneth A. De Jong. *Genetic Algorithms Are NOT Function Optimizers*. 1992.

[106] Kenneth A. De Jong. "Genetic algorithms: A 10 Year Perspective". In: *Proceedings of the 1st International Conference on Genetic Algorithms*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1985, pp. 169–177.

[107] D.A. DeAngelis and G.W. Schulze. "Optimizing piezoelectric ceramic thickness in ultrasonic transducers". In: *Ultrasonic Industry Association Symposium (UIA), 2010 39th Annual*. 2010, pp. 1–9.

[108] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE Transactions on Evolutionary Computation* 6.2 (2002), pp. 182–197.

[109] Mike Deeth. "Nuclear fusion power plant having a liquid reactor core of molten glass that is made laseractive and functions as a tritium breeding blanket which is capable of acousticly compressing/confining fuel so that it radiates and triggers outgoing laser cascades that will reflect from the blast chamber's spherical inside wall and return like photonic Tsunamis, crushing, heating, and causing thermonuclear ignition of the fuel so that heat engines and piezoelectric harvesters can convert the released energy into electricity". Patent US/2012/014491 (A1). Jan. 2012.

[110] J. H. Degnan, G. P. Baca, D. E. Bell, G. Bird, A. L. Chelsey, S. K. Coffey, M. E. Dearborn, M. R. Douglas, S. E. Englert, T. J. Englert, D. Gale, J. D. Graham, K. E. Hackett, J. H. Holmes, T. W. Hussey, G. F. Kiuttu, F. M. Lehr, G. J. Marklin, B. W. Mullins, R. E. Peterkin, D. W. Price, N. F. Roderick, E. L. Ruden, M. Scott, S. W. Seiler, W. Sommars, and P. J. Turchi. "Compression of compact toroids in conical-coaxial geometry". In: *Fusion Science and Technology* 27.2 (Mar. 1995), pp. 107–114.

[111] J. H. Degnan, R. E. Peterkin Jr, G. P. Baca, J. D. Beason, D. E. Bell, M. E. Dearborn, D. Dietz, M. R. Douglas, S. E. Englert, T. J. Englert, K. E. Hackett, J. H. Holmes, T. W. Hussey, G. F. Kiuttu, F. M. Lehr, G. J. Marklin, B. W. Mullins, D. W. Price, N. F. Roderick, E. L. Ruden, C. R. Sovinec, P. J. Turchi, G. Bird, S. K. Coffey, S. W. Seiler, Y. G. Chen, D. Gale, J. D. Graham, M. Scott, and W. Sommars. "Compact toroid formation, compression, and acceleration". In: *Physics of Fluids B: Plasma Physics (1989-1993)* 5.8 (Aug. 1993), pp. 2938–2958.

[112] J.H. Degnan, D.J. Amdahl, A. Brown, T. Cavazos, S.K. Coffey, M.T. Domonkos, M.H. Frese, S.D. Frese, D.G. Gale, T.C. Grabowski, T.P. Intrator, R.C. Kirkpatrick, G.F. Kiuttu, F.M. Lehr, J.D. Letterio, J.V. Parker, R.E. Peterkin, N.F. Roderick, E.L. Ruden, R.E. Siemon, W. Sommars, W. Tucker, P.J. Turchi, and G.A. Wurden. "Experimental and Computational Progress on Liner Implosions for Compression of FRCs". In: *IEEE Transactions on Plasma Science* 36.1 (Feb. 2008), pp. 80–91.

[113] Damián Dellavale, Ludmila Rechiman, Juan Manuel Rosselló, and Fabián Bonetto. "Upscaling energy concentration in multifrequency single-bubble sonoluminescence with strongly degassed sulfuric acid". In: *Physical Review E* 86.1 (July 2012), p. 016320.

[114] Joaquín Derrac, Salvador García, Daniel Molina, and Francisco Herrera. "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms". In: *Swarm and Evolutionary Computation* 1.1 (Mar. 2011), pp. 3–18.

[115] Bishwajyoti Dey and Serge Aubry. "New suggestion concerning the origin of sonoluminescence". In: *Physica D: Nonlinear Phenomena* 216.1 (Apr. 2006), pp. 136–156.

[116] George Djorgovski. "Lecture Ay 20: Basic Astronomy and the Galaxy". `http://www.astro.caltech.edu/~george/ay20/Ay20-Lec7x.pdf`. (visited on 2017/04/23). 2004.

[117] Marco Dorigo, Mauro Birattari, and Thomas Stützle. "Metaheuristic". In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Springer US, Jan. 2010, pp. 662–662.

[118] *DURAN properties*. `http://www.duran-group.com/en/about-duran/duran-properties.html`. (visited on 2014/06/04).

[119] Werner Ebeling, Ingo Rechenberg, Hans-Paul Schwefel, and Hans-Michael Voigt. *Parallel Problem Solving from Nature - PPSN IV: International Conference on Evolutionary Computation. The 4th International Conference on Parallel Problem Solving from Nature, Berlin, Germany, September 22 - 26, 1996. Proceedings*. Springer, Sept. 1996.

[120] R. Eberhart and J. Kennedy. "A new optimizer using particle swarm theory". In: *, Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995. MHS '95.* 1995, pp. 39–43.

[121] Claudia Eberlein. "Theory of quantum radiation observed as sonoluminescence". In: *Physical Review A* 53.4 (Apr. 1996), pp. 2772–2787.

[122] ITER Physics Basis Editors, ITER Physics Expert Group Chairs and Co-Chairs, and ITER Joint Central Team and Physics Integration Unit. "Chapter 1: Overview and summary". In: *Nuclear Fusion* 39.12 (Dec. 1999), p. 2137.

[123] Niles Eldredge and Stephen Jay Gould. "Punctuated equilibria: an alternative to phyletic gradualism". In: *Models in paleobiology*. San Francisco: Cooper & Co, 1972.

[124] Ehab Z. Elfeky, Ruhul A. Sarker, and Daryl L. Essam. "Analyzing the Simple Ranking and Selection Process for Constrained Evolutionary Optimization". In: *Journal of Computer Science and Technology* 23.1 (Jan. 2008), pp. 19–34.

[125] R. C. Elton. "Atomic Processes". In: *Volume 9, Plasma Physics*. Ed. by Hans R. Griem and Ralph H. Lovberg. Vol. 9. Methods in Experimental Physics. Academic Press, 1971, pp. 115–168.

[126] Mark J. Embrechts, Richard T. Lahey Jr., and Robert I. Nigmatulin. "A nonperiodically forced bubble fusion reactor". Patent WO/1996/021230. July 1996.

[127] A. G. Es'kov, R. Kh. Kurtmullaev, A. I. Malyutin, V. N. Semenov, and Y. A. Shipuk. "Liner compression of a toroidal high-β plasma". In: *Proceedings of the third topical conference on pulsed high beta plasmas*. Culham Laboratory, Abingdon, Oxfordshire, UK, 1976, pp. 489–492.

[128] *European Fusion Development Agreement*. `http://www.efda.org/jet/jet%E2%80%99s-main-features/jets-specifications/`. (visited on 2014/02/06).

[129] Emiliano Feresin. "Italian scientists win battle to halt controversial research". In: *Nature News* (June 2012).

[130] C. Fernandes and A. Rosa. "A study on non-random mating and varying population size in genetic algorithms using a royal road function". In: *Proceedings of the 2001 Congress on Evolutionary Computation, 2001*. Vol. 1. IEEE, 2001, 60–66 vol. 1.

[131] M. P. Fewell. "The atomic nuclide with the highest mean binding energy". In: *American Journal of Physics* 63 (July 1995), pp. 653–658.

[132] Frank Boring Fitzgerald. "Method of generating electrical and heat energies via controlled and fail-safe fusion of deuterium in D2O bubbles cycled in radius from energies of ultra-sonic sound and amplitude modulated UHF EM in a narrow liquid D2O reaction gap between a pair of transducers and reactor therefore". Patent US/2009/0022256 A1. Jan. 2009.

[133] David J. Flannigan and Kenneth S. Suslick. "Inertially confined plasma in an imploding bubble". In: *Nature Physics* 6.8 (2010), pp. 598–601.

[134] David J. Flannigan and Kenneth S. Suslick. "Plasma formation and temperature measurement during single-bubble cavitation". In: *Nature* 434.7029 (Mar. 2005), pp. 52–55.

[135] Hugh G. Flynn. "Method of generating energy by acoustically induced cavitation fusion and reactor therefor". Patent US/4333796 (A). June 1982.

[136] Hugh G. Flynn. "Method of generating energy by acoustically induced cavitation fusion and reactor therefor". In: *The Journal of the Acoustical Society of America* 73.2 (1983), pp. 713–713.

[137] D. B Fogel. "Nils Barricelli - artificial life, coevolution, self-adaptation". In: *IEEE Computational Intelligence Magazine* 1.1 (Feb. 2006), pp. 41–45.

[138] D.B. Fogel and R.W. Anderson. "Revisiting Bremermann's genetic algorithm. I. Simultaneous mutation of all parameters". In: *Proceedings of the 2000 Congress on Evolutionary Computation, 2000*. Vol. 2. 2000, 1204–1209 vol.2.

[139] David B. Fogel. *Evolutionary Computation: The Fossil Record*. 1st ed. Wiley-IEEE Press, May 1998.

[140] David B. Fogel. "Revisiting Overlooked Foundations of Evolutionary Computation: Part I". In: *Cybernetics and Systems* 41.5 (2010), pp. 343–358.

[141] David B. Fogel. "Revisiting Overlooked Foundations of Evolutionary Computation: Part II". In: *Cybernetics and Systems* 41.6 (2010), pp. 407–415.

[142] Edward Forringer. "Affidavit of Edward Forringer". `http://newenergytimes.com/v2/bubblegate/Aff/ForringerAffidavit.pdf`. (visited on 2016/04/24). Jan. 2008.

[143] Edward R. Forringer, David Robbins, and Jonathan Martin. "Confirmation of Neutron Production During Self-Nucleated Acoustic Cavitation". In: *Transactions of the American Nuclear Socierty* 95.1 (Nov. 2006), pp. 736–737.

[144] A. S. Fraser. "Simulation of genetic systems". In: *Journal of Theoretical Biology* 2.3 (May 1962), pp. 329–346.

[145] A. S. Fraser. "Simulation of Genetic Systems by Automatic Digital Computers I. Introduction". In: *Aust. Jnl. Of Bio. Sci.* 10.4 (Jan. 1957), pp. 484–491.

[146] Jeffrey P. Freidberg. *Plasma Physics and Fusion Energy*. Cambridge University Press, 2007.

[147] H. Frenzel and H. Schultes. "Luminescenz im ultraschallbeschickten Wasser". In: *Z. Phys. Chem. B* 27 (1934), pp. 421–424.

[148] Lothar Frommhold and Anthony A. Atchley. "Is Sonoluminescence due to Collision-Induced Emission?" In: *Physical Review Letters* 73.21 (Nov. 1994), pp. 2883–2886.

[149] Wojciech Fundamenski. *Power Exhaust in Fusion Plasmas*. Cambridge University Press, Jan. 2010.

[150] D. Felipe Gaitan, Lawrence A. Crum, Charles C. Church, and Ronald A. Roy. "Sonoluminescence and bubble dynamics for a single, stable, cavitation bubble". In: *The Journal of the Acoustical Society of America* 91.6 (1992), pp. 3166–3183.

[151] D. Felipe Gaitan, Yuri A. Pishchalnikov, Thomas J. Matula, Charles C. Church, Joel Gutierrez, Corey Scott, R. Glynn Holt, and Lawrence A. Crum. "Transient cavitation in high-quality factor resonators at high static pressures." In: *The Journal of the Acoustical Society of America* 129.4 (Apr. 2011), pp. 2619–2619.

[152] Dario Felipe Gaitan. “An experimental investigation of acoustic cavitation in gaseous liquids”. doctoral thesis. Lafayette, MI: University of Mississippi, 1990.

[153] Aaron Galonsky. “Tabletop Fusion Revisited”. In: *Science* 297.5587 (June 2002), pp. 1645–1645.

[154] S. F. Garanin. “The MAGO system (magnetic compression)”. In: *IEEE Transactions on Plasma Science* 26.4 (1998), pp. 1230–1238.

[155] S.F. Garanin, V.I. Mamyshev, and V.B. Yakubov. “The MAGO System: Current Status”. In: *IEEE Transactions on Plasma Science* 34.5 (2006), pp. 2273–2278.

[156] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1990.

[157] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1979.

[158] L. Gaul. “The influence of damping on waves and vibrations”. In: *Mechanical Systems and Signal Processing* 13.1 (Jan. 1999), pp. 1–30.

[159] M. Gaviano and D. Lera. “Complexity of general continuous minimization problems: a survey”. In: *Optimization Methods and Software* 20.4-5 (2005), pp. 525–544.

[160] Henry Gee. “Table-top nuclear fusion”. In: *Nature News* (Aug. 1999).

[161] R. Geisler, W.-D. Schmidt-Ott, T. Kurz, and W. Lauterborn. “Search for neutron emission in laser-induced cavitation”. In: *EPL (Europhysics Letters)* 66.3 (May 2004), p. 435.

[162] Mivhel Gendreau and Jean-Ives Potvin, eds. *Handbook of Metaheuristics*. 2nd ed. International Series in Operations Research & Management Science. New York Dordrecht Heidelberg London: Springer, 2010.

[163] *General Fusion Website*. `http://www.generalfusion.com/`. (visited on 2017/11/28). Burnaby, BC.

[164] Josiah Willard Gibbs. “On the equilibrium of heterogeneous substances : first [-second] part”. In: *Transactions of the Connecticut Academy of Arts and Sciences* III.2 (1874).

[165] D. A. Glaser. “Progress report on the development of bubble chambers”. In: *Il Nuovo Cimento Series 9* 11.2 (Jan. 1954), pp. 361–368.

[166] D.A. Glaser. “Invention of the bubble chamber and subsequent events”. In: *Nuclear Physics B - Proceedings Supplements* 36 (July 1994), pp. 3–18.

[167] Donald A. Glaser. “Some Effects of Ionizing Radiation on the Formation of Bubbles in Liquids”. In: *Physical Review* 87.4 (Aug. 1952), pp. 665–665.

[168] Fred Glover. “A Template for Scatter Search and Path Relinking”. In: 1363 (1998), pp. 13–54.

[169] Fred Glover. "Future paths for integer programming and links to artificial intelligence". In: *Computers & Operations Research* 13.5 (1986), pp. 533–549.

[170] Fred Glover. "Genetic algorithms and scatter search: unsuspected potentials". In: *Statistics and Computing* 4.2 (June 1994), pp. 131–140.

[171] Fred Glover. "Heuristics for Integer Programming Using Surrogate Constraints". In: *Decision Sciences* 8.1 (Jan. 1977), pp. 156–166.

[172] Fred Glover, Manuel Laguna, and Rafael Martí. "Scatter search". In: *Advances in Evolutionary Computing: Theory and Applications*. Springer-Verlag, 2003, pp. 519–537.

[173] David Goldberg, Kelsey Milman, and Christina Tidd. *Genetic Algorithms: A Bibliography*. Tech. rep. 1992.

[174] B. Gompf, R. Günther, G. Nick, R. Pecha, and W. Eisenmenger. "Resolving Sonoluminescence Pulse Width with Time-Correlated Single Photon Counting". In: *Physical Review Letters* 79.7 (Aug. 1997), pp. 1405–1408.

[175] Barbara Goss Levi. "Skepticism greets claim of bubble fusion". In: *Physics Today* 55.4 (Apr. 2002).

[176] K. Gottstein. "Die Blasenkammer und ihre Anwendung in der Physik der Elementarteilchen". In: *Naturwissenschaften* 46.3 (Jan. 1959), pp. 97–102.

[177] Stephen Jay Gould and Niles Eldredge. "Punctuated Equilibria: The Tempo and Mode of Evolution Reconsidered". In: *Paleobiology* 3.2 (1977), pp. 115–151.

[178] Andrei A. Goverdovskii, Vladimir S. Imshennik, and Valentin P. Smirnov. "On the prospects of bubble cavitation-induced fusion". In: *Physics-Uspekhi* 56.4 (Apr. 2013), p. 423.

[179] Ed Grayzeck. *Sun Fact Sheet*. `http://nssdc.gsfc.nasa.gov/planetary/factsheet/sunfact.html`. (visited on 2013/04/03). 2012.

[180] Harvey P. Greenspan and Ali Nadim. "On sonoluminescence of an oscillating gas bubble". In: *Physics of Fluids A: Fluid Dynamics* 5.4 (Apr. 1993), pp. 1065–1067.

[181] André Gsponer and Jean-Pierre Hurni. *The physical principles of thermonuclear explosives, inertial confinement fusion, and the quest for fourth generation nuclear weapons*. Tech. rep. Geneva, Switzerland: Independent Scientific Research Institute, Jan. 2009.

[182] Yibin Gu and G.H. Miley. "Experimental study of potential structure in a spherical IEC fusion device". In: *IEEE Transactions on Plasma Science* 28.1 (Feb. 2000), pp. 331–346.

[183] Erico Guizzo. *Bubble Fusion Research Under Scrutiny*. (visited on 2016/03/09). May 2006.

[184] Alexander V. Gurevich and Kirill P. Zybin. "Runaway Breakdown and the Mysteries of Lightning". In: *Physics Today* 58.5 (2005), pp. 37–43.

[185] Walter J. Gutjahr. “Stochastic Search in Metaheuristics”. In: *Handbook of Metaheuristics*. Ed. by Michel Gendreau and Jean-Yves Potvin. International Series in Operations Research & Management Science 146. Springer US, Jan. 2010, pp. 573–597.

[186] Dominik Hammer and Lothar Frommhold. “Light emission of sonoluminescent bubbles containing a rare gas and water vapor”. In: *Physical Review E* 65.4 (Apr. 2002), p. 046309.

[187] Dominik Hammer and Lothar Frommhold. “Sonoluminescence: How bubbles glow”. In: *Journal of Modern Optics* 48.2 (2001), pp. 239–277.

[188] Dominik Hammer and Lothar Frommhold. “Spectra of Sonoluminescent Rare-Gas Bubbles”. In: *Physical Review Letters* 85.6 (Aug. 2000), pp. 1326–1329.

[189] Ming Han. “Diaphragm-based extrinsic Fabry-Perot interferometric optical fiber sensor for acoustic wave detection under high background pressure”. In: *Optical Engineering* 44.6 (June 2005), p. 060506.

[190] Martin Hanke. *personal corresponding*. Cadfem GmbH, Berlin, Germany, 2014.

[191] Nikolaus Hansen. *CMA Evolution Strategy Source Code*. `https://www.lri.fr/%7Ehansen/cmaes_inmatlab.html`. (visited on 2013/10/17).

[192] Nikolaus Hansen. *COmparing Continuous Optimisers: COCO*. `http://coco.gforge.inria.fr/doku.php`. INRIA Centre de Recherche, Saclay, Île-de-France. (visited on 2013/10/04).

[193] Nikolaus Hansen. *Compilation of Results on the 2005 CEC Benchmark Function Set*. `http://web.mysites.ntu.edu.sg/epnsugan/PublicSite/Shared%20Documents/CEC2005/compareresults.pdf`. (visited on 2011/06/10). Institute of Computational Science, ETH Zürich, May 2006.

[194] Nikolaus Hansen, Anne Auger, Steffen Finck, and Raymond Ros. *Real-Parameter Black-Box Optimization Benchmarking: Experimental Setup*. Tech. rep. Saclay, Île-de-France: INRIA, Université Paris Sud, Apr. 2013.

[195] Nikolaus Hansen, Steffen Finck, Raymond Ros, and Anne Auger. *Real-Parameter Black-Box Optimization Benchmarking 2013: Noiseless Functions Definitions*. Tech. rep. RR-6829. Saclay, Île-de-France: INRIA, Université Paris Sud, Apr. 2013.

[196] Nikolaus Hansen and Andreas Ostermeier. “Completely derandomized self-adaptation in evolution strategies”. In: *EVOLUTIONARY COMPUTATION* 9 (2001), pp. 159–195.

[197] Nikolaus Hansen, Andreas Ostermeier, and Andreas Gawelczyk. “On the Adaptation of Arbitrary Normal Mutation Distributions in Evolution Strategies: The Generating Set Adaptation”. In: (1995), pp. 312–317.

[198] William E. Hart and Richard K. Belew. “Optimizing an Arbitrary Function is Hard for the Genetic Algorithm”. In: *Proceedings of the Fourth International Conference on Genetic Algorithms*. Morgan Kaufmann, 1991, pp. 190–195.

[199] E. Newton. Harvey. "Sonoluminescence and Sonic Chemiluminescence". In: *Journal of the American Chemical Society* 61.9 (Sept. 1939), pp. 2392–2398.

[200] Susan Hassler and Erico Guizzo. *Purdue Inquiry Clears Bubble Fusion Researcher*. (visited on 2016/02/03). Feb. 2007.

[201] F. Herrera, M. Lozano, and D. Molina. "Continuous scatter search: An analysis of the integration of some combination methods and improvement strategies". In: *European Journal of Operational Research* 169.2 (Mar. 2006), pp. 450–476.

[202] F. Herrera, M. Lozano, and A. M. Sánchez. "A taxonomy for the crossover operator for real-coded genetic algorithms: An experimental study". In: *International Journal of Intelligent Systems* 18.3 (Mar. 2003), pp. 309–338.

[203] F. Herrera, M. Lozano, and J. L. Verdegay. "Tackling Real-Coded Genetic Algorithms: Operators and Tools for Behavioural Analysis". In: *Artificial Intelligence Review* 12.4 (Aug. 1998), pp. 265–319.

[204] Sascha Hilgenfeldt, Siegfried Grossmann, and Detlef Lohse. "A simple explanation of light emission in sonoluminescence". In: *Nature* 398.6726 (Apr. 1999), pp. 402–405.

[205] Sascha Hilgenfeldt, Siegfried Grossmann, and Detlef Lohse. "Sonoluminescence light emission". In: *Physics of Fluids* 11.6 (1999), p. 1318.

[206] Robert A. Hiller, Seth J. Putterman, and Keith R. Weninger. "Time-Resolved Spectra of Sonoluminescence". In: *Physical Review Letters* 80.5 (Feb. 1998), pp. 1090–1093.

[207] M. Hirsch, J. Baldzuhn, C. Beidler, R. Brakel, R. Burhenn, A. Dinklage, H. Ehmler, M. Endler, V. Erckmann, Y. Feng, J. Geiger, L. Giannone, G. Grieger, P. Grigull, H.-J. Hartfuß, D. Hartmann, R. Jaenicke, R. König, H. P. Laqua, H. Maaßberg, K. McCormick, F. Sardei, E. Speth, U. Stroth, F. Wagner, A. Weller, A. Werner, H. Wobig, S. Zoletnik, and W7-AS Team. "Major results from the stellarator Wendelstein 7-AS". In: *Plasma Physics and Controlled Fusion* 50.5 (May 2008), p. 053001.

[208] Robert L. Hirsch. "Inertial-Electrostatic Confinement of Ionized Fusion Gases". In: *Journal of Applied Physics* 38.11 (Oct. 1967), pp. 4522–4534.

[209] John Henry Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. first published in 1975. MIT Press, 1992.

[210] Stephen Howard, Michel Laberge, Lon McIlwraith, Doug Richardson, and James Gregson. "Development of Merged Compact Toroids for Use as a Magnetized Target Fusion Plasma". In: *Journal of Fusion Energy* 28.2 (Nov. 2008), pp. 156–161.

[211] O. A. Hurricane, D. A. Callahan, D. T. Casey, P. M. Celliers, C. Cerjan, E. L. Dewald, T. R. Dittrich, T. Döppner, D. E. Hinkel, L. F. Berzak Hopkins, J. L. Kline, S. Le Pape, T. Ma, A. G. MacPhee, J. L. Milovich, A. Pak, H.-S. Park, P. K. Patel, B. A. Remington, J. D. Salmonson, P. T. Springer, and R. Tommasini. "Fuel gain exceeding unity in an inertially confined fusion implosion". In: *Nature* advance online publication (Feb. 2014).

[212] Lars Magnus Hvattum, Abraham Duarte, Fred Glover, and Rafael Martí. "Designing effective improvement methods for scatter search: an experimental study on global optimization". In: *Soft Computing* 17.1 (Jan. 2013), pp. 49–62.

[213] "IEEE Standard Definitions and Methods of Measurement for Piezoelectric Vibrators". In: *IEEE Std No.177* (1966), pp. 1–19.

[214] "IEEE Standard on Piezoelectricity". In: *ANSI/IEEE Std 176-1987* (1988).

[215] *ILLIAC*. `http://en.wikipedia.org/w/index.php?title=ILLIAC&oldid=573010734`. Page Version ID: 573010734 (visited on 2013/09/24). Sept. 2013.

[216] J. Inaudi and J. Kelly. "Linear Hysteretic Damping and the Hilbert Transform". In: *Journal of Engineering Mechanics* 121.5 (1995), pp. 626–632.

[217] H Ing, R. A Noulty, and T. D McLean. "Bubble detectors – a maturing technology". In: *Radiation Measurements* 27.1 (Feb. 1997), pp. 1–11.

[218] M. Ito, H. Kume, and K. Oba. "Computer Analysis of the Timing Properties in Micro Channel Plate Photomultiplier Tubes". In: *IEEE Transactions on Nuclear Science* 31.1 (Feb. 1984), pp. 408–412.

[219] Thomas Jansen. *Analyzing Evolutionary Algorithms*. Natural Computing Series 4190. Berlin, Heidelberg: Springer, 2013.

[220] Robert Allen Janssen, John Glen Ahles, Thomas David Ehlert, John Gavin Macdonald, Earl C. Jr Mccraw, Patrick Sean Mcnichols, Paul Warren Rasmussen, and Steve Roffers. "Ultrasonic Treatment Chamber for Initiating Thermonuclear Fusion". Patent WO/2009/072063. June 2009.

[221] Peter Jarman. "Sonoluminescence: A Discussion". In: *The Journal of the Acoustical Society of America* 32.11 (1960), pp. 1459–1462.

[222] *JET - The Joint European Torus - a European Success Story*. Tech. rep. European Fusion Development Agreement (EFDA).

[223] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. `http://www.scipy.org/`. (visited on 2017/11/23). 2001–.

[224] Kenneth A. De Jong. "Genetic Algorithms are NOT Function Optimizers". In: *Proceedings of the Second Workshop on Foundations of Genetic Algorithms. Vail, Colorado, USA, July 26-29 1992*. 1992, pp. 5–17.

[225] Stuart A. Kauffman. *At Home in the Universe: The Search for Laws of Self-organization and Complexity*. Oxford University Press, 1995.

[226] Stuart A. Kauffman. “Emergent properties in random complex automata”. In: *Physica D: Nonlinear Phenomena* 10.1-2 (Jan. 1984), pp. 145–156.

[227] Joseph B. Keller and Michael Miksis. “Bubble oscillations of large amplitude”. In: *The Journal of the Acoustical Society of America* 68.2 (Aug. 1980), pp. 628–633.

[228] Graham Kendall, Andrew Parkes, and Kristian Spoerer. *A Survey of NP-Complete Puzzles*. 2008.

[229] Donald Kennedy. “To Publish or Not to Publish”. In: *Science* 295.5561 (Mar. 2002), pp. 1793–1793.

[230] J. Kennedy and R. Eberhart. “Particle swarm optimization”. In: *, IEEE International Conference on Neural Networks, 1995. Proceedings*. Vol. 4. 1995, 1942–1948 vol.4.

[231] James Kennedy. “Particle Swarm Optimization”. In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Springer US, Jan. 2010, pp. 760–766.

[232] Shahzad Khalid, Brian Kappus, Keith Weninger, and Seth Putterman. “Opacity and Transport Measurements Reveal That Dilute Plasma Models of Sonoluminescence Are Not Valid”. In: *Physical Review Letters* 108.10 (Mar. 2012), p. 104302.

[233] A. L. Kimball and D. E. Lovell. “Internal Friction in Solids”. In: *Physical Review* 30.6 (Dec. 1927), pp. 948–959.

[234] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. “Optimization by simulated annealing”. In: *Science* 220 (1983), pp. 671–680.

[235] H. Kita, I. Ono, and S. Kobayashi. “Theoretical analysis of the unimodal normal distribution crossover for real-coded genetic algorithms”. In: *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*. May 1998, pp. 529–534.

[236] J. C. von Vaupel Klein. “Punctuated equilibria and phyletic gradualism: Even partners can be good friends”. In: *Acta Biotheoretica* 42.1 (Mar. 1994), pp. 15–48.

[237] Natalio Krasnogor. “Studies on the Theory and Design Space of Memetic Algorithms”. PhD thesis. University of the West of England at Bristol, 2002.

[238] Natalio Krasnogor and Pier Luca Lanzi, eds. *13th Annual Genetic and Evolutionary Computation Conference, GECCO 2011, Proceedings, Dublin, Ireland, July 12-16, 2011*. ACM, 2011.

[239] Akshata Krishnamurthy. “Development and characterization of an inertial electrostatic confinement thruster”. master thesis. Urbana, IL: University of Illinois at Urbana-Champaign, 2012.

[240] Steven Krivit. “Federal Investigations Reveal Academic Backstabbing at Purdue University”. In: *New Energy Times* (Jan. 2014).

[241] Steven B. Krivit. *New Energy Times - Bubblegate Portal*. http://newenergytimes.com/v2/bubblegate/BubblegatePortal.shtml. (visited on 2013/03/25). 2006.

[242] R.A. Krohling. "Gaussian swarm: a novel particle swarm optimization algorithm". In: *2004 IEEE Conference on Cybernetics and Intelligent Systems*. Vol. 1. 2004, 372–376 vol.1.

[243] Karl S. Kruszelnicki. *Lazy Sun is less energetic than compost*. http://www.abc.net.au/science/articles/2012/04/17/3478276.htm. ABC Science. (visited on 2014/02/07). Apr. 2012.

[244] D. Kuhl and G. Meschke. *Finite Element Methods in Linear Structural Mechanics (course script)*. Bochum: Ruhr University Bochum, 2005.

[245] Kenji Kurakata and Tazu Mizunami. "Comparison of equal-loudness-level contours between otologically normal young and older adults". In: *Acoustical Science and Technology* 35.5 (2014), pp. 243–250.

[246] Heinrich Kuttruff. "Licht aus Schall". In: *Physik in unserer Zeit* 30.1 (1999), pp. 23–30.

[247] Michel Laberge. "An Acoustically Driven Magnetized Target Fusion Reactor". In: *Journal of Fusion Energy* 27.1-2 (July 2007), pp. 65–68.

[248] Gaudencio A. Labrador. "Heat energy recapture and recycle and its new applications". Patent US/2005/120715 (A1). June 2005.

[249] Manuel Laguna, Rafael Martí, and Rafael Cunquero Martí. *Scatter search: methodology and implementations in C*. Springer, 2003.

[250] R.T. Lahey Jr., Y. Danon, Frank J. Saglime III, R.C. Block, F. Morage, S. Cancelos, and M. Kitto. *Rensselaer's Bubble Fusion Project – Final Report*. Tech. rep. Rensselaer Polytechnic Institute, Troy, NY, 2004.

[251] R.T. Lahey Jr., R.P. Taleyarkhan, R.I. Nigmatulin, and I.S. Akhatov. "Sonoluminescence and the Search for Sonofusion". In: *Advances in Heat Transfer*. Ed. by James P. Hartnett George A. Greene. Vol. Volume 39. Elsevier, 2006, pp. 1–168.

[252] Richard T. Lahey Jr. *NYSERDA Progress Report*. Tech. rep. Rennselaer Polytechnic Institute, May 2004.

[253] Richard T. Lahey Jr. *personal corresponding*. 2007-2017.

[254] Richard T. Lahey Jr., Rusi P. Taleyarkhan, and Robert I. Nigmatulin. "Sonofusion technology revisited". In: *Nuclear Engineering and Design* 237.15-17 (Sept. 2007), pp. 1571–1585.

[255] Richard T. Lahey Jr, Rusi P. Taleyarkhan, and Robert I. Nigmatulin. *Bubble Power*. http://spectrum.ieee.org/energy/nuclear/bubble-power/1. (Blog: IEEE Spectrum, visited on 2016/01/16). May 2005.

[256] *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*. NS III/16. Springer.

[257] Werner Lauterborn and Thomas Kurz. “Physics of bubble oscillations”. In: *Reports on Progress in Physics* 73.10 (Oct. 2010), p. 106501.

[258] Henning Leidecker. *Elastic Moduli and Damping of Vibrational Modes of Aluminum/Silicon Carbide Composite Beams*. Technical Report NASA-TM-104626. Greenbelt, MD: NASA Goddard Space Flight Center, Mar. 1996.

[259] T. G. Leighton. *Derivation of the Rayleigh-Plesset Equation in Terms of Volume*. ISVR Technical Report 308. University of Southampton, 2007.

[260] T. G. Leighton. *The Acoustic Bubble*. Elsevier, 1994.

[261] T. Lepoint, D. De Pauw, F. Lepoint-Mullie, M. Goldman, and A. Goldman. “Picosecond sonoluminescence: An alternative electrohydrodynamic hypothesis”. In: *, 12th International Conference on Conduction and Breakdown in Dielectric Liquids, 1996, ICDL ’96*. 1996, pp. 103–106.

[262] Reinhard Lerch, Gerhard M. Sessler, and Dietrich Wolf. *Technische Akustik: Grundlagen und Anwendungen*. Springer, 2009.

[263] C.-Z. Li, H.-Y. Zhang, Y.-C. Chang, W.-P. Zhang, and Y. Liu. “Dielectric breakdown of PZT ferroelectric ceramics”. In: *Second International Conference on Properties and Applications of Properties and Applications of Dielectric Materials, Proceedings*. 1988, 198–201 vol.1.

[264] Jing J. Liang. “Novel Particle Swarm Optimizers with Hybrid, Dynamic & Adaptive Neighborhood Structures”. doctoral thesis. Singapore: Nanyang Technological University, School of Electrical & Electronic Engineering, 2008.

[265] Jing J. Liang, B. Y. Qu, Ponnuthurai Nagaratnam Suganthan, and Alfredo G. Hernández-Díaz. *Problem Definitions and Evaluation Criteria for the CEC 2013 Special Session on Real-Parameter Optimization*. Tech. rep. 201212. Singapore: Nanyang Technological University, Jan. 2013.

[266] Tianjun Liao and Thomas Stützle. “Benchmark results for a simple hybrid algorithm on the CEC 2013 benchmark set for real-parameter optimization”. In: *2013 IEEE Congress on Evolutionary Computation (CEC)*. 2013, pp. 1938–1944.

[267] Bruce S. Lieberman and Niles Eldredge. “What is punctuated equilibrium? What is macroevolution? A response to Pennell et al.” In: *Trends in Ecology & Evolution* 29.4 (Apr. 2014), pp. 185–186.

[268] Haiko Lietz. *Bubble Fusion takes next hurdle*. `http://www.heise.de/tp/artikel/20/20542/1.html`. (Blog: Telepolis, visited on 2013/07/30). July 2005.

[269] Ron Lifshitz. “Phonon-mediated dissipation in micro- and nano-mechanical systems”. In: *Physica B: Condensed Matter* (2002), pp. 397–399.

[270] I. R. Lindemuth, R. E. Reinovsky, R. E. Chrien, J. M. Christian, C. A. Ekdahl, J. H. Goforth, R. C. Haight, G. Idzorek, N. S. King, R. C. Kirkpatrick, R. E. Larson, G. L. Morgan, B. W. Olinger, H. Oona, P. T. Sheehey, J. S. Shlachter, R. C. Smith, L. R. Veeser, B. J. Warthen, S. M. Younger, V. K. Chernyshev, V. N. Mokhov, A. N. Demin, Y. N. Dolin, S. F. Garanin, V. A. Ivanov, V. P. Korchagin, O. D. Mikhailov, I. V. Morozov, S. V. Pak, E. S. Pavlovskii, N. Y. Seleznev, A. N. Skobelev, G. I. Volkov, and V. A. Yakubov. "Target Plasma Formation for Magnetic Compression/Magnetized Target Fusion". In: *Physical Review Letters* 75.10 (Sept. 1995), pp. 1953–1956.

[271] Hermann Linder. *Linder Biologie: Lehrbuch für die Oberstufe*. Schroedel, 2005.

[272] A. G. Lipson, V. A. Klyuev, B. V. Deryagin, Yu. P. Toporov, M. G. Sirotyuk, O. B. Khavroshkin, and D. M. Sakov. "Observation of neutrons accompanying cavitation in deuterium-containing media". In: *Pis'ma Zh. Tekh. Fiz.* 16 (1990), pp. 89–93.

[273] Andrei G. Lipson. "Comment on "Nuclear Emissions During Self-Nucleated Acoustic Cavitation"". In: *Physical Review Letters* 97.14 (Oct. 2006), p. 149401.

[274] Shui-Yin Lo. "A method for generating nuclear fusion through high pressure". Patent WO/1997/049274 A2. Dec. 1997.

[275] Ritva Löfstedt, Bradley P. Barber, and Seth Putterman. "Scaling laws for sonoluminescence". In: *The Journal of the Acoustical Society of America* 92.4 (Oct. 1992), pp. 2453–2453.

[276] Ritva Löfstedt, Bradley P. Barber, and Seth J. Putterman. "Toward a hydrodynamic theory of sonoluminescence". In: *Physics of Fluids A: Fluid Dynamics* 5.11 (Nov. 1993), pp. 2911–2928.

[277] Manuel López-Ibañez, Jérémie Dubois-Lacoste, Leslie Pérez Cáceres, Thomas Stützle, Mauro Birattari, Eric Yuan, and Prasanna Balaprakash. *The irace Package, Iterated Race for Automatic Algorithm Configuration*. `http://iridia.ulb.ac.be/irace/`. (visited on 2014/11/12).

[278] Manuel López-Ibáñez, Jérémie Dubois-Lacoste, Thomas Stützle, and Mauro Birattari. *The irace package, Iterated Race for Automatic Algorithm Configuration*. Tech. rep. TR/IRIDIA/2011-004. IRIDIA, Université Libre de Bruxelles, Belgium, 2011.

[279] M. Lozano, D. Molina, C. García-Martinez, and Herrera. *Evolutionary Algorithms and other Metaheuristics for Continuous Optimization Problems*. `http://sci2s.ugr.es/eamhco/`. (visited on 2013/10/17).

[280] Sean Luke. *Essentials of Metaheuristics*. second edition. Lulu, 2013.

[281] Naresh Mahamuni. "Fuel Loading of Gaseous Fuel in Liquid Metal Cavitation Reactors". Patent US/2012/0312381 A1. Dec. 2012.

[282] Naresh Mahamuni and Peter L. Nelson. "Flow-Through Structure with Active Ingredients". Patent US/2012/0141344 A1. June 2012.

[283] Vittorio Maniezzo, Thomas Stützle, and Stefan Voß. *Matheuristics – Hybridizing Metaheuristics and Mathematical Programming*. Springer, 2010.

[284] Milia A. Margulis. "Sonoluminescence". In: *Physics-Uspekhi* 43.3 (Mar. 2000), pp. 259–282.

[285] Milia A. Margulis and Igor M. Margulis. "Contemporary review on nature of sonoluminescence and sonochemical reactions". In: *Ultrasonics Sonochemistry* 9.1 (Jan. 2002), pp. 1–10.

[286] N. Marinesco and J.-J. Trillat. "Action des ultrasons sur les plaques photographiques". In: *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 196 (1933), pp. 858–860.

[287] Rafael Martí, Manuel Laguna, and Fred Glover. "Principles of scatter search". In: *European Journal of Operational Research* 169.2 (Mar. 2006), pp. 359–372.

[288] G.E. Martin. "Dielectric, Elastic and Piezoelectric Losses in Piezoelectric Materials". In: *1974 Ultrasonics Symposium*. 1974, pp. 613–617.

[289] Mark A. Maslin, Chris M. Brierley, Alice M. Milner, Susanne Shultz, Martin H. Trauth, and Katy E. Wilson. "East African climate pulses and early human evolution". In: *Quaternary Science Reviews* 101 (Oct. 2014), pp. 1–17.

[290] Bernard de Massy. "Distribution of meiotic recombination sites". In: *Trends in Genetics* 19.9 (Sept. 2003), pp. 514–522.

[291] Thomas Matula, Brian MacConnaghy, Lawrence Crum, and Felipe Gaitan. "Comparison of single bubble collapse and cluster collapse in a high pressure vessel." In: *The Journal of the Acoustical Society of America* 129.4 (Apr. 2011), pp. 2619–2619.

[292] Ernst Mayr. *Change of Genetic Environment and Evolution*. London: George Allen & Unwin Limited, 1954.

[293] Ernst Mayr. "Ecological Factors in Speciation". In: *Evolution* 1.4 (1947), pp. 263–288.

[294] Thomas John McGuire. "Improved lifetimes and synchronization behavior in multi-grid inertial electrostatic confinement fusion devices". doctoral thesis. Massachusetts Institute of Technology, Feb. 2007.

[295] Lisa Marie Meffert. "Population Bottlenecks and Founder Effects". In: *Evolutionary Genetics: Concepts and Case Studies*. Ed. by Charles W. Fox and Jason Wolf. Oxford University Press, 2006, p. 608.

[296] Gregor Mendel. "Versuche über Pflanzen-Hybriden". In: *Zeitschrift für Theoretische und Angewandte Genetik (formerly: Der Züchter)* 13.10-11 (Oct. 1941), pp. 221–268.

[297] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (June 1953), p. 1087.

[298] Robert Mettin. "Bubble structures in acoustic cavitation". In: *Bubble and Particle Dynamics in Acoustic Fields: Modern Trends and Applications*. Ed. by Alexander A. Doinikov. Research Signpost, 2006.

[299] Zbigniew Michalewicz. "A perspective on evolutionary computation". In: *Progress in Evolutionary Computation*. Ed. by G. Goos, J. Hartmanis, J. Leeuwen, Jaime G. Carbonell, Jörg Siekmann, and Xin Yao. Vol. 956. Springer Berlin Heidelberg, 1995, pp. 73–89.

[300] George H. Miley and Murali S. Krupakar. *Inertial Electrostatic Confinement (IEC) Fusion – Fundamentals and Applications*. Springer, 2014.

[301] R.L. Miller and R.A. Krakowski. *Assessment of the slowly-imploding liner (LINUS) fusion reactor concept*. Tech. rep. LA-UR-80-3071. King of Prussia, PA: Los Alamos Scientific Laboratory, Jan. 1980.

[302] Kimball A. Milton. "The Casimir effect: recent controversies and progress". In: *Journal of Physics A: Mathematical and General* 37.38 (2004), R209.

[303] Kenro Miyamoto. *Plasma Physics and Controlled Nuclear Fusion*. Springer, Mar. 2006.

[304] Jan Mlynář. *Focus On: JET*. Tech. rep. EFD-R(07)01. European Fusion Development Agreement (EFDA).

[305] *Model 482A21 Installation and Operating Manual*. Document no. 21354. PCB Piezotronics, Inc. 3425 Walden Ave, Depew, NY 14043 USA.

[306] N. P. Moloi and M. M. Ali. "An Iterative Global Optimization Algorithm for Potential Energy Minimization". In: *Computational Optimization and Applications* 30.2 (Feb. 2005), pp. 119–132.

[307] *Monte Carlo N-Particle Transport Code System (MCNP5/MCNPX)*. `https://mcnp.lanl.gov/`. (visited on 2017/11/30). Los Alamos, NM, USA.

[308] M. J. Moran, R. E. Haigh, M. E. Lowry, D. R. Sweider, G. R. Abel, J. T. Carlson, S. D. Lewia, A. A. Atchley, D. F. Gaitan, and X. K. Maruyama. "Direct observations of single sonoluminescence pulses". In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*. The Interaction of Swift Particles and Electromagnetic Fields with Matter 96.3-4 (May 1995), pp. 651–656.

[309] Pablo Moscato. *On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts – Towards Memetic Algorithms*. 1989.

[310] William C. Moss, Douglas B. Clarke, John W. White, and David A. Young. "Hydrodynamic simulations of bubble collapse and picosecond sonoluminescence". In: *Physics of Fluids* 6.9 (Sept. 1994), pp. 2979–2985.

[311] William C. Moss, Douglas B. Clarke, John W. White, and David A. Young. "Sonoluminescence and the prospects for table-top micro-thermonuclear fusion". In: *Physics Letters A* 211.2 (Feb. 1996), pp. 69–74.

[312] William C. Moss, Douglas B. Clarke, and David A. Young. "Calculated Pulse Widths and Spectra of a Single Sonoluminescing Bubble". In: *Science* 276.5317 (May 1997), pp. 1398–1401.

[313] Thomas Most and Johannes Will. "Metamodel of Optimal Prognosis – An automated approach for variable reduction and optimal metamodel selection". In: *Weimar Optimization and Stochastic Days 2008*. Weimar, Germany, 2008.

[314] Heinz Mühlenbein and Dirk Schlierkamp-Voosen. "Predictive Models for the Breeder Genetic Algorithm – I. Continuous Parameter Optimization". In: *Evolutionary Computation* 1 (1993), pp. 25–49.

[315] G.B. Muravskii. "On frequency independent damping". In: *Journal of Sound and Vibration* 274.3-5 (July 2004), pp. 653–668.

[316] Colin Murray. *An Experiment to Save The World*. Ed. by Matthew Barrett, Jonathan Renouf, Orly Danon, and Andy Fegen. BBC documentation series "Horizon". 2005.

[317] Colin Murray. *An Experiment to Save The World – programme transcript*. Ed. by Matthew Barrett, Jonathan Renouf, Orly Danon, and Andy Fegen. `http://www.bbc.co.uk/sn/tvradio/programmes/horizon/experiment_trans.shtml`. (visited on 2016/01/30). 2005.

[318] B. Naranjo. "Comment on "Taleyarkhan et al. Reply:"" in: *arXiv:physics/0702009 [physics.gen-ph]* (Feb. 2007). arXiv: physics/0702009 (visited on 2016/03/21).

[319] B. Naranjo. "Comment on "Nuclear Emissions During Self-Nucleated Acoustic Cavitation"". In: *Physical Review Letters* 97.14 (Oct. 2006), p. 149403.

[320] B. Naranjo, J. K. Gimzewski, and S. Putterman. "Observation of nuclear fusion driven by a pyroelectric crystal". In: *Nature* 434.7037 (Apr. 2005), pp. 1115–1117.

[321] Ahid D. Nashif, David I. G. Jones, and John P. Henderson. *Vibration damping*. 3rd ed. John Wiley & Sons, Inc., 1985.

[322] Richard Neifeld. "Tabletop nuclear fusion generator". Patent US/2007/0002996 A1. Jan. 2007.

[323] J. A. Nelder and R. Mead. "A Simplex Method for Function Minimization". In: *The Computer Journal* 7.4 (Jan. 1965), pp. 308–313.

[324] R. I. Nigmatulin, R. T. Lahey Jr., R. P. Taleyarkhan, C. D. West, and R. C. Block. "On thermonuclear processes in cavitation bubbles". In: *Physics-Uspekhi* 57.9 (Sept. 2014), pp. 877–890.

[325] R. I. Nigmatulin, R. P. Taleyarkhan, and R. T. Lahey Jr. "Evidence for nuclear emissions during acoustic cavitation revisited". In: *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy* 218.5 (Aug. 2004), pp. 345–364.

[326] Robert I. Nigmatulin, Iskander Sh. Akhatov, Andrey S. Topolnikov, Raisa Kh. Bolotnova, Nailya K. Vakhitova, Richard T. Lahey Jr., and Rusi P. Taleyarkhan. "Theory of supercompression of vapor bubbles and nanoscale thermonuclear fusion". In: *Physics of Fluids* 17.10 (2005), p. 107106.

[327] B. E. Noltingk and E. A. Neppiras. "Cavitation produced by Ultrasonics". In: *Proceedings of the Physical Society. Section B* 63.9 (1950), p. 674.

[328] Patrik Nosil, Jeffrey L. Feder, Samuel M. Flaxman, and Zachariah Gompert. "Tipping points in the dynamics of speciation". In: *Nature Ecology & Evolution* 1.2 (Jan. 2017), s41559–016–0001–016.

[329] Lukas Novotny. "Strong coupling, energy splitting, and level crossings: A classical perspective". In: *American Journal of Physics* 78.11 (Oct. 2010), pp. 1199–1202.

[330] A. S. Nowick and B. S. Berry, eds. *Anelastic Relaxation in Crystalline Solids*. Materials Science Series. Academic Press, 1972.

[331] Danail Obreschkow, Philippe Kobel, Aurèle De Bosset, Nicolas Dorsaz, Claude Nicollier, and Mohamed Farhat. *Des bulles, des bulles! L'expérience à bien fonctionné*. `http://letemps.blogs.com/apesanteur/2006/03/des_bulles_des_.html`. (visited on 2014/09/28). Mar. 2006.

[332] Y.-S. Ong, N. Krasnogor, and H. Ishibuchi. "Special Issue on Memetic Algorithms". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 37.1 (Feb. 2007), pp. 2–5.

[333] Yew-Soon Ong, Meng-Hiot Lim, Ferrante Neri, and Hisao Ishibuchi. "Special issue on emerging trends in soft computing: memetic algorithms". In: *Soft Computing* 13 (July 2008), pp. 739–740.

[334] Isao Ono, Hajime Kita, and Shigenobu Kobayashi. "A Robust Real-coded Genetic Algorithm Using Unimodal Normal Distribution Crossover Augmented by Uniform Crossover: Effects of Self-adaptation of Crossover Probabilities". In: *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation – Volume 1*. GECCO'99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 496–503.

[335] Isao Ono and Shigenobu Kobayashi. "A Real-coded Genetic Algorithm for Function Optimization Using Unimodal Normal Distribution Crossover". In: *Proceedings of the 7th International Conference on Genetic Algorithms ICGA-97* (1997), pp. 346–253.

[336] Y. Ono, M. Inomoto, Y. Ueda, T. Matsuyama, and T. Okazaki. "New relaxation of merging spheromaks to a field reversed configuration". In: *Nuclear Fusion* 39.11Y (Nov. 1999), p. 2001.

[337] *OpenOpt*. `http://openopt.org/Welcome`. (visited on 2013/10/17).

[338] Ibrahim H. Osman and Gilbert Laporte. "Metaheuristics: A bibliography". In: *Annals of Operations Research* 63.5 (Oct. 1996), pp. 511–623.

[339] Craig Packer and Anne E. Pusey. "Adaptations of Female Lions to Infanticide by Incoming Males". In: *The American Naturalist* 121.5 (May 1983), pp. 716–728.

[340] José Antonio Parejo, Antonio Ruiz-Cortés, Sebastián Lozano, and Pablo Fernandez. "Metaheuristic optimization frameworks: a survey and benchmarking". In: *Soft Computing* 16.3 (Mar. 2012), pp. 527–561.

[341] J. Park, R. A. Nebel, S. Stange, and S. Krupakar Murali. "Experimental Observation of a Periodically Oscillating Plasma Sphere in a Gridded Inertial Electrostatic Confinement Device". In: *Physical Review Letters* 95.1 (June 2005), p. 015003.

[342] Jaeyoung Park, Nicholas A. Krall, Paul E. Sieck, Dustin T. Offermann, Michael Skillicorn, Andrew Sanchez, Kevin Davis, Eric Alderson, and Giovanni Lapenta. "High Energy Electron Confinement in a Magnetic Cusp Configuration". In: *arXiv:1406.0133 [physics.plasm-ph]* (June 2014). (visited on 2015/03/09).

[343] K. E. Parsopoulos and M. N. Vrahatis. "Recent approaches to global optimization problems through Particle Swarm Optimization". In: *Natural Computing* 1.2-3 (June 2002), pp. 235–306.

[344] Hermanus Pauw. "Energy spectra of radioactive neutron sources". doctoral thesis. Amsterdam, Netherlands: University of Amsterdam, 1970.

[345] R. Pecha and B. Gompf. "Microimplosions: Cavitation Collapse and Shock Wave Emission on a Nanosecond Time Scale". In: *Physical Review Letters* 84.6 (Feb. 2000), pp. 1328–1330.

[346] Mark Peplow. "Desktop fusion is back on the table". In: *Nature News* (Oct. 2006).

[347] Susanne Pfalzner. *An Introduction to Inertial Confinement Fusion*. CRC Press, Mar. 2006.

[348] Daniel Phillips, Ross Tessien, and Richard Satterwhite. "Hourglass-shaped cavitation chamber". Patent US/2006/0269429 A1. Nov. 2006.

[349] Daniel Phillips, Ross Tessien, and Richard Satterwhite. "Hourglass-shaped cavitation chamber with spherical lobes". Patent US/2006/0269430 A1. Nov. 2006.

[350] *Photomultiplier Tubes – Construction and Operating Characteristics Connections to External Circuits*. Hamamatsu Photonics, K. K. 314-5, Shimokanzo, Toyooka-village, Iwata-gun, Shizuoka-ken, 438-0193, Japan, 1998.

[351] Alexander Piel. *Plasma Physics: An Introduction to Laboratory, Space, and Fusion Plasmas*. Springer, June 2010.

[352] *Piezoelectric ceramic material and measurements guidelines for sonar transducers*. military standard MIL-STD-1376B. US Department of Defence, Feb. 1995.

[353] *Piezoelectric Ceramics, Catalog*. Channel Industries, Inc. 839 Ward Drive, Santa Barbara, CA 93111, USA, 2008.

[354] Yuri A. Pishchalnikov, Joel Gutierrez, Wylene W. Dunbar, and Richard W. Philpott. "Intense cavitation at extreme static pressure". In: *Ultrasonics* 65 (Feb. 2016), pp. 380–389.

[355] Irwin A. Pless. "Method and apparatus for generating large velocity, high pressure, and high temperature conditions". Patent US/5968323 (A). Oct. 1999.

[356] D. R. Poelman. *On the science of lightning: an overview*. Tech. rep. 56. Koninklijk Meteorologisch Instituut van België, 2010.

[357] Riccardo Poli, James Kennedy, and Tim Blackwell. "Particle swarm optimization". In: *Swarm Intelligence* 1.1 (June 2007), pp. 33–57.

[358] G. Polya. *How to Solve It: A New Aspect of Mathematical Method*. 2nd ed. Princeton University Press, 1957.

[359] Steve Poterala. *personal corresponding*. Channel Technologies Group, Santa Barbara, CA, 2014.

[360] D.J. Powell, J. Mould, and G.L. Wojcik. "Dielectric and mechanical absorption mechanisms for time and frequency domain transducer modeling". In: *Proceedings of the IEEE Ultrasonics Symposium 1998*. Vol. 2. 1998, pp. 1019–1024.

[361] M. J. D. Powell. "An efficient method for finding the minimum of a function of several variables without calculating derivatives". In: *The Computer Journal* 7.2 (Jan. 1964), pp. 155–162.

[362] Kenneth V. Price, Rainer M. Storn, and Jouni A. Lampinen. *Differential evolution: a practical approach to global optimization*. Birkhäuser, 2005.

[363] *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2005, 2-4 September 2005, Edinburgh, UK*. IEEE, 2005.

[364] *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2011, New Orleans, LA, USA, 5-8 June, 2011*. IEEE, 2011.

[365] Addy Pross. "Toward a general theory of evolution: Extending Darwinian theory to inanimate matter". In: *Journal of Systems Chemistry* 2.1 (June 2011), p. 1.

[366] William B. Provine. "Ernst Mayr: Genetics and Speciation". In: *Genetics* 167.3 (July 2004), pp. 1041–1046.

[367] S. J. Putterman, L. A. Crum, and K. Suslick. "Comments on "Evidence for Nuclear Emissions During Acoustic Cavitation" by R.P. Taleyarkhan et al., Science volume 295,p.1868, March 8, 2002". In: *arXiv:cond-mat/0204065 [cond-mat.soft]* (Apr. 2002). (visited on 2016/03/19).

[368] S. J. Putterman and K. R. Weninger. "Sonoluminescence: How Bubbles Turn Sound into Light". In: *Annual Review of Fluid Mechanics* 32.1 (Jan. 2000), pp. 445–476.

[369] Seth Putterman. "Sonoluminescence: the star in a jar". In: *Physics World* 11.May 1998 (May 1998), pp. 38–42.

[370] Seth J. Putterman. "Sonoluminescence: Sound into Light". In: *Scientific American* 272.2 (Feb. 1995), pp. 46–51.

[371] Seth J. Putterman, Bradley Paul Barber, Robert Anthony Hiller, and Ritva Maire Johanna Löfstedt. "Converting acoustic energy into useful other energy forms". Patent US/5659173 A. Aug. 1997.

[372] Seth Putterman, Bradley Barber, Robert Hiller, and Ritva Löfstedt. "Converting Acoustic Energy into Useful Other Energy Forms". Patent WO/1995/023413. Sept. 1995.

[373] *PyGMO 1.1.5 documentation*. `http://pagmo.sourceforge.net/pygmo/`. (visited on 2013/10/17).

[374] A. K. Qin, V. L. Huang, and P. N. Suganthan. "Differential Evolution Algorithm With Strategy Adaptation for Global Numerical Optimization". In: *IEEE Transactions on Evolutionary Computation* 13.2 (Apr. 2009), pp. 398–417.

[375] Daniel J. Rader, Wayne M. Trott, John R. Torczynski, Jaime N. Castañeda, and T.W. Grasser. *Measurement of Thermal Accomodation Coefficients*. Sandia Report SAND2005-6084. Albuquerque, New Mexico 87185 and Livermore, California 94550: Sandia National Laboratories, Oct. 2005.

[376] Thomas Rapp. *Rapp Instruments*. `http://www.rapp-instruments.de`. (visited on 2015/03/03). Liesel-Beckmann-Straße 2, D-81369 München.

[377] Gregory J. E. Rawlins, ed. *Foundations of genetic algorithms*. Elsevier Science & Tech, 1991.

[378] Ingo Rechenberg. *Evolutionsstrategie Computer-Animationen*. `http://www.bionik.tu-berlin.de/institut/s2anima.html`. (visited on 2013/03/14).

[379] Ingo Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog, 1973.

[380] Ingo Rechenberg. *Vorlesungen: Bionik und Evolutionsstrategie*. `http://www.bionik.tu-berlin.de/institut/skript/s2skript.htm`. (visited on 2013/09/04). 2011.

[381] Ludmila M. Rechiman, Fabián J. Bonetto, and Juan M. Rosselló. "Effect of the Rayleigh-Taylor instability on maximum reachable temperatures in laser-induced bubbles". In: *Physical Review E* 86.2 (Aug. 2012), p. 027301.

[382] Jon Reed, Robert Toombs, and Nils Aall Barricelli. "Simulation of biological evolution and machine learning : I. Selection of self-reproducing numeric patterns by data processing machines, effects of hereditary control, mutation type and crossing". In: *Journal of Theoretical Biology* 17.3 (Dec. 1967), pp. 319–342.

[383] Eugenie Samuel Reich. "Bubble fusion: silencing the hype". In: *Nature News* (Mar. 2006).

[384] Eugenie Samuel Reich. "Is bubble fusion simply hot air?" In: *Nature News* (Mar. 2006).

[385] Eugenie Samuel Reich. "Purdue attacked over fusion inquiry". In: *Nature* 444.7120 (Dec. 2006), pp. 664–665.

[386] D. Ress, L. B. DaSilva, R. A. London, J. E. Trebes, S. Mrowka, R. J. Procassini, T. W. Barbee, and D. E. Lehr. "Measurement of Laser-Plasma Electron Density with a Soft X-ray Laser Deflectometer". In: *Science* 265.5171 (July 1994). PMID: 17781311, pp. 514–517.

[387] D: Richardson, A. Froese, V. Suponitsky, M. Reynolds, and D. Plant. "Status of Progress Towards Acoustic Magnetized Target Fusion at General Fusion". In: Toronto, ON, June 2013.

[388] Dirk Roos, Ulrike Adam, and Christian Bucher. "Robust Design Optimization". In: *3rd Weimar Optimization and Stochastic Days*. Weimar, Germany, Nov. 2006.

[389] Grzegorz Rozenberg. *Handbook of Natural Computing*. Springer, Sept. 2011.

[390] Frank J. Saglime III. "Experimental Results for the RPI Bubble Fusion Project". master thesis. Troy, NY: Rensselaer Polytechnic Institute, 2004.

[391] Frank J. Saglime III. "High energy neutron differential scattering measurements for beryllium and molybdenum". doctoral thesis. Troy, NY: Rensselaer Polytechnic Institute, 2009.

[392] Frank J. Saglime III. *personal corresponding*. 2007.

[393] Frank J. Saglime III. *personal corresponding*. 2007-2017.

[394] Ralf Salomon. "Reevaluating Genetic Algorithm Performance under Coordinate Rotation of Benchmark Functions - A survey of some theoretical and practical aspects of genetic algorithms". In: *BIOSYSTEMS* 39 (1995), pp. 263–278.

[395] M. J. Saltmarsh and Dan Shapira. "Questions Regarding Nuclear Emissions in Cavitation Experiments". In: *Science* 297.5587 (June 2002), pp. 1603–1603.

[396] Ernesto Salzano and Anna Basco. "Comparison of the Explosion Thermodynamics of TNT and Black Powder Using Le Chatelier Diagrams". In: *Propellants, Explosives, Pyrotechnics* 37.6 (2012), pp. 724–731.

[397] Ian Sample. "Science runs into trouble with bubbles". In: *The Guardian* (Mar. 2004).

[398] Wallapat Santawisuk, Widchaya Kanchanavasita, Chakrit Sirisinha, and Choltacha Harnirattisai. "Dynamic viscoelastic properties of experimental silicone soft lining materials". In: *Dental Materials Journal* 29.4 (2010), pp. 454–460.

[399] Richard D. Satterwhite. "High pressure cavitation chamber with dual internal reflectors". Patent US/7510322 B2. Mar. 2009.

[400] Richard D. Satterwhite. "Method of operating a high pressure cavitation chamber with dual internal reflectors". Patent US/7461966 (B2). Dec. 2008.

[401] W. Schaaffs. *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*. Ed. by K.-H. Hellwege and A. M. Hellwege. NS II/5. Springer, 1967.

[402] M. M. Schauer, D. C. Barnes, and K. R. Umstadter. "Physics of non-thermal Penning-trap electron plasma and application to ion trapping". In: *Physics of Plasmas (1994-present)* 11.1 (Jan. 2004), pp. 9–15.

[403] Tom Schaul, Justin Bayer, Daan Wierstra, Yi Sun, Martin Felder, Frank Sehnke, Thomas Rückstieß, and Jürgen Schmidhuber. *PyBrain – The python modern machine learning library*. `http://www.pybrain.org/pages/home`. code project. (visited on 2012/11/07).

[404] *Schott Duran® Borosilikat 3.3 Laborglas*. Duran Produktions GmbH & Co. KG. Hattenbergstraße 10, 55122 Mainz, Germany.

[405] C. Schumacher, M. D. Vose, and L. D. Whitley. "The No Free Lunch and Problem Description Length". In: *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*. Morgan Kaufmann, 2001, pp. 565–570.

[406] Hans-Paul Schwefel. "Direct search for optimal parameters within simulation models". In: *Proceedings of the 12th annual symposium on Simulation*. ANSS '79. Piscataway, NJ, USA: IEEE Press, 1979, pp. 91–102.

[407] Hans-Paul Schwefel. *Numerical optimization of computer models*. New York, NY: Wiley, 1981.

[408] J. Schwinger. In: *Proceedings of the National Academy of Sciences* 90 (1993), pp. 958, 2105, 4505, 7285.

[409] Charles Seife. "'Bubble Fusion' Paper Generates a Tempest in a Beaker". In: *Science* 295.5561 (Mar. 2002), pp. 1808–1809.

[410] Charles Seife. *Sun in a bottle: the strange history of fusion and the science of wishful thinking*. New York: Viking, 2008.

[411] Frederick Seitz. "On the Theory of the Bubble Chamber". In: *Physics of Fluids (1958-1988)* 1.1 (1958), pp. 2–13.

[412] Aldo M. Serenelli, Sarbani Basu, Jason W. Ferguson, and Martin Asplund. "New Solar Composition: The Problem with Solar Models Revisited". In: *The Astrophysical Journal Letters* 705.2 (2009), p. L123.

[413] S. Seth, M. Das, S. Bhattacharya, P. Bhattacharjee, and S. Saha. "The nucleation parameter for heavy-ion induced bubble nucleation in superheated emulsion detector". In: *Journal of Instrumentation* 8.05 (May 2013), P05001.

[414] D. Shapira and M. Saltmarsh. "Nuclear Fusion in Collapsing Bubbles – Is It There? An Attempt to Repeat the Observation of Nuclear Emissions from Sonoluminescence". In: *Physical Review Letters* 89.10 (Aug. 2002), p. 104302.

[415] D. Shapira and M. J. Saltmarsh. *Comments on the possible observation of d-d fusion in sonoluminescence*. Tech. rep. Oak Ridge, TN: Physics Division, Oak Ridge National Laboratory, Feb. 2002.

[416] Dan Shapira and Mike Saltmarsh. "Nuclear fusion in collapsing bubbles – Is it there? An attempt to repeat an experiment that reported d-d fusion in bubble collapse induced by cavitation in deuterated acetone". In: *The Journal of the Acoustical Society of America* 113.4 (Apr. 2003), p. 2223.

[417] S. Sherrit, H. D. Wiederick, B. K. Mukherjee, and M. Sayer. "An accurate equivalent circuit for the unloaded piezoelectric vibrator in the thickness mode". In: *Journal of Physics D: Applied Physics* 30.16 (Aug. 1997), p. 2354.

[418] S. Sherrit, H. D. Wiederick, and B.K. Mukherjee. "Accurate equivalent circuits for unloaded piezoelectric resonators". In: *, 1997 IEEE Ultrasonics Symposium, 1997. Proceedings.* Vol. 2. Oct. 1997, 931–935 vol.2.

[419] A. R. Sherwood, B. L. Freeman, R. A. Gerwin, T. R. Jarboe, R. A. Krakowski, R. C. Malone, J. Marshall, R. L. Miller, and B. Suydam. *Fast liner proposal.* Tech. rep. LA-6707-P. Los Alamos, NM: Los Alamos Scientific Laboratory, 1977.

[420] Kalachandra Sid and Tetsuya Takamata. "Polymers in the Oral Environments". In: *Biomaterials Engineering and Devices: Human Applications – Volume 2. Orthopedic, Dental, and Bone.* Vol. 2. Springer, 2000.

[421] Andreas Skyman. *Skyman's licensiate thesis.* `https://gitlab.com/skyman-s-licensiate-thesis/skyman-s-licensiate-thesis`. (visited on 2017/02/04). 2011.

[422] Andreas Skyman. "Turbulent impurity transport in tokamak fusion plasmas". Licentiate of Engineering. Göteborg: Chalmers University of Technology, 2011.

[423] Boris M. Smirnov. *Fundamentals of Ionized Gases.* 1st ed. Berlin: Wiley-VCH, Nov. 2011.

[424] Francisco J. Solis and Roger J.-B. Wets. "Minimization by Random Search Techniques". In: *Mathematics of Operations Research* 6.1 (Feb. 1981), pp. 19–30.

[425] Weston M. Stacey. *Fusion: An Introduction to the Physics and Technology of Magnetic Confinement Fusion.* John Wiley & Sons, Feb. 2010.

[426] D. Stansfield. *Underwater electroacoustic transducers – A handbook for users and designers.* Bath; St. Albans: Bath University Press; Institute of Acoustics, 1991.

[427] Michael Stix. "On the time scale of energy transport in the sun". In: *Solar Physics* 212.1 (Jan. 2003), pp. 3–6.

[428] M. J. Stokmaier, A. G. Class, and T. Schulenberg. "A hard optimisation test function with symbolic solution visualisation for fast interpretation by the human eye". In: *2013 IEEE Congress on Evolutionary Computation (CEC).* 2013, pp. 2251–2258.

[429] M. J. Stokmaier, A. G. Class, T. Schulenberg, and R. T. Lahey Jr. "Optimising acoustic resonators for sonofusion experiments with evolutionary algorithms". In: *Proceedings of the International Conference on the Applications of Computer Science and Mathematics in Architecture and Civil Engineering.* Bauhaus-University Weimar, July 2015.

[430] Markus J. Stokmaier. *peabox – an evolutionary algorithm toolbox written in python.* `https://github.com/stromatolith/peabox`. (visited on 2013/03/17). Sept. 2012.

[431] Markus J. Stokmaier. *RP_Bubble – a minimal Python script for simulating the Rayleigh-Plesset equation*. `https://github.com/stromatolith/RP_Bubble`. (visited on 2017/12/03). Mar. 2017.

[432] Markus J. Stokmaier. *zycircle – python scripts for analysing admittance and impedance circles*. `https://github.com/stromatolith/zycircle`. (visited on 2014/04/04). Mar. 2014.

[433] Markus J. Stokmaier, Andreas G. Class, Thomas Schulenberg, and Richard T. Lahey Jr. "FE modelling of vibration behaviour of acoustic resonator for sonofusion experiments". In: *Jahrestagung Kerntechnik*. Berlin, Germany, May 2010.

[434] Markus J. Stokmaier, Andreas G. Class, Thomas Schulenberg, and Richard T. Lahey Jr. "Sonofusion: EA optimisation of acoustic resonator". In: *PAMM* 12.1 (2012), pp. 623–624.

[435] Markus J. Stokmaier, Andreas G. Class, Thomas Schulenberg, Richard T. Lahey Jr., and Bernard A. Malouin. "On the operating characteristics of acoustic chambers for sonofusion". In: *NURETH-13: Proceedings of the 13th international topical meeting on nuclear reactor thermal hydraulics*. Kanazawa, Ishikawa (Japan), Oct. 2009.

[436] Markus J. Stokmaier, Richard T. Lahey Jr., Andreas G. Class, Bernard A. Malouin, and Thomas Schulenberg. "Acoustic chambers for sonofusion experiments – sensitivity on geometry and materials". In: *Bulletin of the American Physical Society*. Vol. Volume 54, Number 19. Minneapolis, Minnesota: American Physical Society, Nov. 2009.

[437] Markus J. Stokmaier, Richard T. Lahey Jr., Andreas G. Class, Thomas Schulenberg, and Bernard A. Malouin. "Acoustic chambers for sonofusion experiments – FE-analysis highlighting performance limiting factors". In: *Proceedings of the 17th international congress on sound and vibration*. Cairo, Egypt, July 2010.

[438] R. Storn and K. Price. "Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces". In: *J. of Global Optimization* 11.4 (1997), pp. 341–359.

[439] Rainer Storn. *Differential Evolution Homepage*. `http://www1.icsi.berkeley.edu/~storn/code.html`. (visited on 2013/10/17).

[440] Rainer Storn and Kenneth Price. *Differential Evolution – A simple and efficient adaptive scheme for global optimization over continuous spaces*. Tech. rep. TR-95-012. 1947 Center Street, Berkeley, CA: International Computer Science Institute, Mar. 1995, pp. 1–12.

[441] Matthew J. Streeter. "Two Broad Classes of Functions for Which a No Free Lunch Result Does Not Hold". In: *Genetic and Evolutionary Computation – GECCO 2003*. Lecture Notes in Computer Science 2724. Springer Berlin Heidelberg, Jan. 2003, pp. 1418–1430.

[442] Roger Stringham. "Cavitation reactor and method of producing heat". Patent WO/2005/028985. Mar. 2005.

[443] P. N. Suganthan. *IEEE Congress on Evolutionary Computation, competition results 2005-2011, hosted on Prof. P. N. Suganthan's homepage*. `http://www.ntu.edu.sg/home/epnsugan/`. (visited on 2012/11/07).

[444] Ponnuthurai Nagaratnam Suganthan, Nikolaus Hansen, Jing J. Liang, Kalyanmoy Deb, Y. P. Chen, Anne Auger, and S. Tiwari. *Problem Definitions and Evaluation Criteria for the CEC 2005 Special Session on Real-Parameter Optimization*. Tech. rep. KanGAL 2005005. Singapore: Nanyang Technological University, 2005.

[445] Jonathan R. Sukovich, Ashwin Sampathkumar, Phillip A. Anderson, R. Glynn Holt, Yuri A. Pishchalnikov, and D. Felipe Gaitan. "Temporally and spatially resolved imaging of laser-nucleated bubble cloud sonoluminescence". In: *Physical Review E* 85.5 (May 2012), p. 056605.

[446] Chang T. Sun and Yeh-Pei Lu. *Vibration damping of structural elements*. 1st ed. Englewood Cliffs, NJ: Prentice Hall PTR, 1995.

[447] Victoria Suponitsky, Aaron Froese, and Sandra Barsky. "Richtmyer-Meshkov instability of a liquid-gas interface driven by a cylindrical imploding pressure wave". In: *Computers & Fluids* 89 (Jan. 2014), pp. 1–19.

[448] Kenneth S. Suslick, Nathan C. Eddingsaas, David J. Flannigan, Stephen D. Hopkins, and Hangxun Xu. "Extreme conditions during multibubble cavitation: Sonoluminescence as a spectroscopic probe". In: *Ultrasonics Sonochemistry* 18.4 (July 2011), pp. 842–846.

[449] Kenneth S. Suslick and David J. Flannigan. "Inside a Collapsing Bubble: Sonoluminescence and the Conditions During Cavitation". In: *Annual Review of Physical Chemistry* 59.1 (May 2008), pp. 659–683.

[450] Gilbert Syswerda. "Uniform Crossover in Genetic Algorithms". In: *Proceedings of the 3rd International Conference on Genetic Algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989, pp. 2–9.

[451] J. M. Taccetti, T. P. Intrator, G. A. Wurden, S. Y. Zhang, R. Aragonez, P. N. Assmus, C. M. Bass, C. Carey, S. A. deVries, W. J. Fienup, I. Furno, S. C. Hsu, M. P. Kozar, M. C. Langner, J. Liang, R. J. Maqueda, R. A. Martinez, P. G. Sanchez, K. F. Schoenberg, K. J. Scott, R. E. Siemon, E. M. Tejero, E. H. Trask, M. Tuszewski, W. J. Waganaar, C. Grabowski, E. L. Ruden, J. H. Degnan, T. Cavazos, D. G. Gale, and W. Sommars. "FRX-L: A field-reversed configuration plasma injector for magnetized target fusion". In: *Review of Scientific Instruments* 74.10 (Oct. 2003), pp. 4314–4323.

[452] El-Ghazali Talbi. *Metaheuristics: From Design to Implementation*. John Wiley & Sons, May 2009.

[453] R. P. Taleyarkhan, R. C. Block, R. T. Lahey Jr., R. I. Nigmatulin, and Y. Xu. "Taleyarkhan et al. Reply:" in: *Physical Review Letters* 97.14 (Oct. 2006), p. 149404.

[454] R. P. Taleyarkhan, R. C. Block, R. T. Lahey Jr., R. I. Nigmatulin, and Y. Xu. "Taleyarkhan et. al Reply:" in: *Physical Review Letters* 97.14 (Oct. 2006), p. 149402.

[455] R. P. Taleyarkhan, R. C. Block, C. D. West, and R: T. Lahey Jr. "Comments on the Shapira and Saltmarsh Report". Oak Ridge National Laboratory, Mar. 2002.

[456] R. P. Taleyarkhan, J. S. Cho, C. D. West, R. T. Lahey Jr., R. I. Nigmatulin, and R. C. Block. "Additional evidence of nuclear emissions during acoustic cavitation". In: *Physical Review E* 69.3 (Mar. 2004), p. 036109.

[457] R. P. Taleyarkhan, Richard T. Lahey Jr., and R. I. Nigmatulin. "Bubble Nuclear Fusion Technology – Status and challenges". In: *Multiphase Science and Technology* 17.3 (2005), pp. 191–224.

[458] R. P. Taleyarkhan, C. D. West, J. S. Cho, R. T. Lahey Jr., R. I. Nigmatulin, and R. C. Block. "Evidence for Nuclear Emissions During Acoustic Cavitation". In: *Science* 295.5561 (Mar. 2002), pp. 1868–1873.

[459] R. P. Taleyarkhan, C. D. West, R. T. Lahey Jr., R. I. Nigmatulin, R. C. Block, and Y. Xu. "Nuclear Emissions During Self-Nucleated Acoustic Cavitation". In: *Physical Review Letters* 96.3 (Jan. 2006), p. 034301.

[460] Rusi Taleyarkhan. "Nanoscale explosive-implosive burst generators using nuclear-mechanical triggering of pretensioned liquids". Patent US/2003/0074010 A1. Apr. 2003.

[461] Rusi P. Taleyarkhan. "Acoustic Inertial Confinement Nuclear Device". Patent WO/2008/013571. Feb. 2008.

[462] Rusi P. Taleyarkhan. "Acoustic inertial confinement nuclear fusion device". Patent US/2010/254500 (A1). Oct. 2010.

[463] Rusi P. Taleyarkhan, Richard T. Lahey, and Robert I. Nigmatulin. "Acoustic Inertial Confinement Nuclear Fusion". In: *Nuclear Energy Encyclopedia*. Ed. by Steven B. Krivit, Jay H. Lehr, and Thomas B. Kingery. John Wiley & Sons, Inc., 2011, pp. 551–567.

[464] Rusi P. Taleyarkhan and Colin D. West. "Methods and apparatus to induce D-D and D-T reactions". Patent US/2005/0135532 A1. June 2005.

[465] Rusi P. Taleyarkhan, Colin D. West, JaeSeon Cho, Richard T. Lahey Jr., Robert I. Nigmatulin, and Robert C. Block. "Comments on Letter (Phys. Rev. L, Vol.89, No. 10,2002) by D. Shapira and M. Saltmarsh". In: *arXiv:1307.3217 [physics.gen-ph]* 1307.3217 (July 2013). (visited on 2013/07/30).

[466] Rusi P. Taleyarkhan, Colin D. West, Richard T. Lahey Jr., Robert I. Nigmatulin, R. C. Block, J. S. Cho, and Y. Xu. "Recent Advances and Results in Acoustic Inertial Confinement Bubble Nuclear Fusion". In: *Low-Energy Nuclear Reactions and New Energy Technologies Sourcebook*. Ed. by Jan Marwan and Steven B. Krivit. Vol. 2. Washington, DC: American Chemical Society, 2009.

[467] Rusi Taleyarkhan and Colin West. "Methods and Apparatus to Induce D-D and D-T Reactions". Patent WO/2002/097823. Dec. 2002.

[468] P. R. Temple, R. W. Detenbeck, and W. L. Nyborg. "Sonoluminescence from a "Shuttlecock" Associated with a Single Gas Bubble". In: *The Journal of the Acoustical Society of America* 50.1A (1971), pp. 112–112.

[469] Alan R. Templeton. “The reality and importance of founder speciation in evolution”. In: *BioEssays* 30.5 (May 2008), pp. 470–479.

[470] Alan R. Templeton. “The Theory of Speciation VIA the Founder Principle”. In: *Genetics* 94.4 (Apr. 1980), pp. 1011–1038.

[471] Ross Tessien. “Cavitation Nuclear Reactor”. Patent WO/2001/039202. June 2001.

[472] Ross Tessien. “Cavitation nuclear reactor”. Patent WO/2001/039203 A2. May 2001.

[473] Ross Tessien. “Shaped Core Cavitation Nuclear Reactor”. Patent WO/2001/039204. June 2001.

[474] Ross Alan Tessien. “Fluid Rotation and Hydraulic Actuation Systems for a Cavitation Chamber”. Patent WO/2006/078434. July 2006.

[475] Ross Alan Tessien. “Magnetic fluid rotation system for a cavitation chamber”. Patent US/8157433 B2. Apr. 2012.

[476] Ross Alan Tessien. “Method of fabricating a spherical cavitation chamber utilizing electron beam welding”. Patent US/7571531 B2. Aug. 2009.

[477] Nicholas A. Tomory. “A Sonofusion Device and Method of Operating the Same”. Patent WO/2007/101340 (A1). Sept. 2007.

[478] Y. Toriyabe, E. Yoshida, J. Kasagi, and M. Fukuhara. “Acceleration of the d+d reaction in metal lithium acoustic cavitation with deuteron bombardment from 30 to 70 keV”. In: *Physical Review C* 85.5 (May 2012), p. 054620.

[479] Nicholas W. Tschoegl. *The phenomenological theory of linear viscoelastic behavior*. Berlin, Heidelberg, New York: Springer, 1989.

[480] Lefteri Tsoukalas, Franklin Clikeman, Martin Bertodano, Tatjana Jevremovic, Joshua Walter, Anton Bougaev, and Edward Merritt. “Tritium Measurements in Neutron-Induced Cavitation of Deuterated Acetone”. In: *Nuclear Technology* 155.2 (Aug. 2006), pp. 248–251.

[481] Shigeyoshi Tsutsui, Masayuki Yamamura, and Takahide Higuchi. “Multi-parent recombination with simplex crossover in real coded genetic algorithms”. In: *Proceedings of the Genetic and Evolutionary Computation (GECCO '99)* (1999), pp. 657–664.

[482] P. J. Turchi, A. L. Cooper, R. D. Ford, D. J. Jenkins, and R. L. Burton. “Review of the NRL Liner Implosion Program”. In: *Megagauss Physics and Technology*. Ed. by Peter J. Turchi. Springer US, 1980, pp. 375–386.

[483] P. J. Turchi, A. L. Cooper, R. D. Ford, D. J. Jenkins, and W. L. Warnick. *Stabilized Liner Implosions Driven by Axially-Moving Free-Piston*. NRL Memorandum Report 3511. Washington, D.C.: Naval Research Laboratory, 1977.

[484] Peter J. Turchi. *Stabilized Liquid Liner Implosions for Repetitive Compression of Plasma Targets*. `https://www.arpa-e.energy.gov/sites/default/files/documents/files/Drivers_Fusion_Turchi_Presentation.pdf`. (visited on 2017/11/23). Berkeley, CA, Oct. 2013.

[485] Erik Vance. "The Bursting of Bubble Fusion". In: *The Chronicle of Higher Education* (Apr. 2007).

[486] Massimiliano Vasile and Paolo DePascale. "On the Preliminary Design of Multiple Gravity-Assist Trajectories". In: *arXiv:1105.1822 [math.OC]* (May 2011). (visited on 2017/04/14).

[487] Massimiliano Vasile, Juan Manuel Romero Martin, Luca Masi, Edmondo Minisci, Richard Epenoy, Vincent Martinot, and Jordi Fontdecaba Baig. "Incremental planning of multi-gravity assist trajectories". In: *Acta Astronautica* 115 (Oct. 2015), pp. 407–421.

[488] Philippe Vaxelaire. "Modular unit for a tubular ultrasonic reactor". Patent US/5384508 (A). Jan. 1995.

[489] G. Vazquez, C. Camara, S. Putterman, and K. Weninger. "Sonoluminescence: Nature's Smallest BlackBody". In: *arXiv:physics/0009057 [physics.flu-dyn]* (Sept. 2000). (visited on 2012/05/09).

[490] Shankar Vedantam. *Fusion Experiment Sparks an Academic Brawl*. `https://www.washingtonpost.com/archive/politics/2002/03/11/fusion-experiment-sparks-an-academic-brawl/f5d03c08-1771-4627-9a25-b2817a64caf5/`. (visited on 2016/04/03). Mar. 2002.

[491] Victor V. Verbinski, Francis G. Perey, J. Kirk Dickens, and Walter R. Burrus. "Neutrons from $^{9}\mathrm{Be}(\alpha, n)$ Reaction for E_α between 6 and 10 MeV". In: *Physical Review* 170.4 (June 1968), pp. 916–923.

[492] Michael D. Vose. *A Critical Examination of the Schema Theorem*. Tech. rep. Knoxville, TN, USA: University of Tennessee, 1993.

[493] Stefan Voß. "Meta-heuristics: The State of the Art". In: *Local Search for Planning and Scheduling*. Ed. by Alexander Nareyek. Lecture Notes in Computer Science 2148. Springer Berlin Heidelberg, Jan. 2001, pp. 1–23.

[494] Michael Wahl. *Time-Correlated Single Photon Counting*. Tech. rep. Rudower Chaussee 29, 12489 Berlin: PicoQuant GmbH, 2014.

[495] Jing Wang. "Numerical simulation and experimental study on resonant acoustic chambers". master thesis. West Lafayette, Indiana: Purdue University, Dec. 2008.

[496] Jing Wang, Brian Archambault, Yiban Xu, and Rusi P. Taleyarkhan. "Numerical simulation and experimental study on Resonant Acoustic Chambers – For novel, high-efficiency nuclear particle detectors". In: *Nuclear Engineering and Design* 240.11 (Nov. 2010), pp. 3716–3726.

[497] Jing Wang, Brian Archambault, Yiban Xu, and Rusi P. Taleyarkhan. "Numerical Simulation and Experimental Study on Resonant Acoustic Chambers: For Novel, High-Efficiency Nuclear Particle Detectors". In: *Proceedings of the 17th international conference on nuclear engineering (ICONE17)*. Brussels, Belgium, July 2009, pp. 623–633.

[498] Jing Wang, Yiban Xu, and Rusi P. Taleyarkhan. "Modeling and benchmarking of resonant acoustic chambers". In: *12th International Meeting on Nuclear Reactor Thermal Hydraulics (NURETH 12)*. Pittsburgh, PA, Sept. 2007, pp. 1290–1303.

[499] Xiang Wang, Hanxing Liu, and Shixi Ouyang. "Damping properties of flexible epoxy resin". In: *Journal of Wuhan University of Technology-Mater. Sci. Ed.* 23.3 (June 2008), pp. 411–414.

[500] A. H. Wapstra, G. Audi, and C. Thibault. "The Ame2003 atomic mass evaluation: (I). Evaluation of input data, adjustment procedures". In: *Nuclear Physics A*. The 2003 NUBASE and Atomic Mass Evaluations 729.1 (Dec. 2003), pp. 129–336.

[501] S. M. Webb and N. J. Mason. "Single-bubble sonoluminescence: creating a star in a jar". In: *European Journal of Physics* 25.1 (Jan. 2004), p. 101.

[502] C. F. V. Weizsäcker. "Zur Theorie der Kernmassen". In: *Zeitschrift fur Physik* 96 (July 1935), pp. 431–458.

[503] John Wesson. *The Science of JET*. JET Report JET-R(99)13. Dec. 1999.

[504] C. West. *Cavitation nucleation by energetic particles*. United Kingdom Atomic Energy Authority Research Group Report AERE - R 5486. Harwell, Berkshire: Atomic Energy Research Establishment (AERE), 1967.

[505] C. West and R. Howlett. "Some experiments on ultrasonic cavitation using a pulsed neutron source". In: *Journal of Physics D: Applied Physics* 1.2 (Feb. 1968), pp. 247–254.

[506] Colin West. "Affidavit of Dr. Colin West". `http://newenergytimes.com/v2/bubblegate/Aff/West.pdf`. (visited on 2016/04/05). Jan. 2008.

[507] Darrell Whitley. "A genetic algorithm tutorial". In: *Statistics and Computing* 4.2 (June 1994), pp. 65–85.

[508] Darrell Whitley. *Genitor – Genetic Algorithms Research at Colorado State*. `http://www.cs.colostate.edu/~genitor/functions.html`. (visited on 2013/02/19).

[509] Darrell Whitley. "The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best". In: *Proceedings of the Third International Conference on Genetic Algorithms*. San Mateo, CA: Morgan Kaufmann, 1989, pp. 116–121.

[510] Darrell Whitley and Monte Lunacek. "Ruffled by ridges: How evolutionary algorithms can fail". In: *Genetic and Evolutionary Computation – GECCO 2004*. Springer Verlag, 2004, pp. 294–306.

[511] Darrell Whitley, Soraya Rana, John Dzubera, and Keith E. Mathias. "Evaluating evolutionary algorithms". In: *Artificial Intelligence* 85.1-2 (Aug. 1996), pp. 245–276.

[512] Darrell Whitley and Jean Paul Watson. "Complexity Theory and the No Free Lunch Theorem". In: *Search Methodologies*. Ed. by Edmund K. Burke and Graham Kendall. Springer US, Jan. 2005, pp. 317–339.

[513] *Wikipedia: evolutionary algorithm*. http://en.wikipedia.org/wiki/Evolutionary_algorithm. Page Version ID: 521285025 (visited on 2012/11/07). Nov. 2012.

[514] *Wikipedia: metaheuristic*. http://en.wikipedia.org/wiki/Metaheuristics. Page Version ID: 520149614 (visited on 2012/11/07). Oct. 2012.

[515] Daniel Nicolas Wilke. "Analysis of the particle swarm optimization algorithm". master thesis. Pretoria: University of Pretoria, Department of Mechanical and Aeronautical Engineering, Feb. 2005.

[516] Oscar Bryan Wilson. *Introduction to theory and design of sonar transducers*. Los Altos, CA: Peninsula, 1988.

[517] Friedwardt Winterberg. *The physical principles of thermonuclear explosive devices*. Fusion energy foundation frontiers of science. New York, NY: Fusion Energy Foundation, 1981.

[518] Friedwardt Winterberg. *The Release of Thermonuclear Energy by Inertial Confinement: Ways Towards Ignition*. World Scientific, 2010.

[519] D.H. Wolpert and W.G. Macready. "No free lunch theorems for optimization". In: *IEEE Transactions on Evolutionary Computation* 1.1 (1997), pp. 67–82.

[520] C. C. Wu and Paul H. Roberts. "Shock-wave propagation in a sonoluminescing gas bubble". In: *Physical Review Letters* 70.22 (May 1993), pp. 3424–3427.

[521] G. A. Wurden, K. F. Schoenberg, R. E. Siemon, M. Tuszewski, F. J. Wysocki, and R. D. Milroy. "Magnetized Target Fusion: a burning FRC plasma in an imploded metal can". In: *Journal of Plasma and Fusion Research SERIES* 2 (1999), pp. 238–241.

[522] Glen A. Wurden. *Realizing Technologies for Magnetized Target Fusion*. http://wsx.lanl.gov/talks/MTF-Reactor-Technologies-TOFE-2012-Wurden-inv-talk.pdf. (visited on 2017/11/23). Nashville, TN, Aug. 2012.

[523] Hongfeng Xiao and Guanzheng Tan. "A Novel Simplex Hybrid Genetic Algorithm". In: *Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference for*. 2008, pp. 1801–1806.

[524] Y. Xu, A. Butt, and S. T. Revankar. "Bubble Dynamics and Tritium Emission During Bubble Fusion Experiments". In: *NURETH-11: Proceedings of the 11th international topical meeting on nuclear reactor thermal hydraulics*. Avignon, France, 2005.

[525] Yiban Xu. "To: Purdue University Committee (Affidavit)". http://newenergytimes.com/v2/bubblegate/Aff/Xu-Yiban-Jan31-2008.pdf. (visited on 2016/04/04). West Lafayette, Indiana, Jan. 2008.

[526] Yiban Xu and Adam Butt. "Confirmatory experiments for nuclear emissions during acoustic cavitation". In: *Nuclear Engineering and Design* 235.10-12 (May 2005). Festschrift Edition Celebrating the 65th Birthday of Prof. Richard T. Lahey Jr., pp. 1317–1324.

[527] Kyuichi Yasui. “Effect of liquid temperature on sonoluminescence”. In: *Physical Review E* 64.1 (June 2001), p. 016310.

[528] Kyuichi Yasui. “Mechanism of single-bubble sonoluminescence”. In: *Physical Review E* 60.2 (Aug. 1999), pp. 1754–1758.

[529] Kyuichi Yasui, Toru Tuziuti, Manickam Sivakumar, and Yasuo Iida. “Sonoluminescence”. In: *Applied Spectroscopy Reviews* 39.3 (2004), pp. 399–436.

[530] Kyuichi Yasui, Toru Tuziuti, Manickam Sivakumar, and Yasuo Iida. “Theoretical study of single-bubble sonochemistry”. In: *The Journal of Chemical Physics* 122.22 (June 2005), pp. 224706–224706–12.

[531] K. Yosioka and A. Omura. “The light emission from a single bubble driven by ultrasound and the spectra of acoustic oscillations”. In: *Proc. Annu. Meet. Acoust. Soc. Jpn* May 1962 (1962).

[532] F. Ronald Young. *Sonoluminescence*. CRC Press, 2004.

[533] Daniela Zaharie. “Influence of crossover on the behavior of Differential Evolution Algorithms”. In: *Applied Soft Computing* 9.3 (June 2009), pp. 1126–1138.

[534] Günther Zäpfel, Roland Braune, and Michael Bögl. *Metaheuristic Search Concepts - A Tutorial with Applications to Production and Logistics*. Springer, 2010.

[535] J. Zhang, R. J. Perez, and E. J. Lavernia. “Documentation of damping capacity of metallic, ceramic and metal-matrix composite materials”. In: *Journal of Materials Science* 28.9 (May 1993), pp. 2395–2404.

[536] Jingqiao Zhang and A. C. Sanderson. “JADE: Adaptive Differential Evolution With Optional External Archive”. In: *IEEE Transactions on Evolutionary Computation* 13.5 (Oct. 2009), pp. 945–958.

[537] “Zickzack nach Darwin”. In: *Der Spiegel* 47 (Nov. 1964), pp. 145–147.

[538] James F. Ziegler. *The Stopping and Range of Ions in Matter (SRIM)*. `http://www.srim.org`. (visited on 2014/02/19).

[539] O. C. Zienkiewicz and R. L. Taylor. *The Finite Element Method, The Basis*. Vol. 1. Oxford: Butterworth-Heinemann, Oct. 2000.

Own publications

[op1] M. Stokmaier, G. Goll, C. Sürgers, D. Weissenberger, F. Pérez-Willard, and H. v. Löhneysen. "Spin-dependent transport through nanostructured S/F point contacts". In: *Verhandlungen der Deutschen Physikalischen Gesellschaft* 41.1 (2006).

[op2] Markus Julius Stokmaier. "Untersuchung des spinpolarisierten Transports durch nanostrukturierte Al/Fe-Punktkontakte". Diplomarbeit. Karlsruhe: Universität Karlsruhe (TH), 2006.

[op3] H. v. Löhneysen, D. Beckmann, G. Goll, F. Pérez Willard, H. Stalzer, M. Stokmaier, and C. Sürgers. "Proximity effect between superconductors and ferromagnets". In: *Physica C: Superconductivity and its Applications*. Proceedings of the 8th International Conference on Materials and Mechanisms of Superconductivity and High Temperature Superconductors 460-462.Part 1 (Sept. 2007), pp. 322–326.

[op4] M. Stokmaier, G. Goll, D. Weissenberger, C. Sürgers, and H. v. Löhneysen. "Size Dependence of Current Spin Polarization through Superconductor/Ferromagnet Nanocontacts". In: *Physical Review Letters* 101.14 (Oct. 2008), p. 147005.

[op5] Markus J. Stokmaier, Andreas G. Class, Thomas Schulenberg, Richard T. Lahey Jr., and Bernard A. Malouin. "On the operating characteristics of acoustic chambers for sonofusion". In: *NURETH-13: Proceedings of the 13th international topical meeting on nuclear reactor thermal hydraulics*. Kanazawa, Ishikawa (Japan), Oct. 2009.

[op6] Markus J. Stokmaier, Richard T. Lahey Jr., Andreas G. Class, Bernard A. Malouin, and Thomas Schulenberg. "Acoustic chambers for sonofusion experiments – sensitivity on geometry and materials". In: *Bulletin of the American Physical Society*. Vol. Volume 54, Number 19. Minneapolis, Minnesota: American Physical Society, Nov. 2009.

[op7] Markus J. Stokmaier, Andreas G. Class, Thomas Schulenberg, and Richard T. Lahey Jr. "FE modelling of vibration behaviour of acoustic resonator for sonofusion experiments". In: *Jahrestagung Kerntechnik*. Berlin, Germany, May 2010.

[op8] Markus J. Stokmaier, Richard T. Lahey Jr., Andreas G. Class, Thomas Schulenberg, and Bernard A. Malouin. "Acoustic chambers for sonofusion experiments – FE-analysis highlighting performance limiting factors". In: *Proceedings of the 17th international congress on sound and vibration*. Cairo, Egypt, July 2010.

[op9] Christoph Sürgers, Ajay Singh, Markus Stokmaier, Gernot Goll, Fabian Pérez-Willard, and Hilbert v. Löhneysen. "Spintronics in metallic superconductor/ferromagnet hybrid structures". In: *International Journal of Materials Research* 101.2 (Feb. 2010), pp. 164–174.

[op10] Markus J. Stokmaier, Andreas G. Class, Thomas Schulenberg, and Richard T. Lahey Jr. "Liquid-filled acoustic resonator for sonofusion experiments – solutions to the design and the algorithmic optimisation problems". In: *ENC 2012 Transactions*. Manchester, UK, Dec. 2012.

[op11] Markus J. Stokmaier, Andreas G. Class, Thomas Schulenberg, and Richard T. Lahey Jr. "Sonofusion: EA optimisation of acoustic resonator". In: *PAMM* 12.1 (2012), pp. 623–624.

[op12] S. Bouvron, M. Stokmaier, M. Marz, and G. Goll. "Andreev experiments on superconductor/ferromagnet point contacts". In: *Low Temperature Physics* 39.3 (Mar. 2013), pp. 274–278.

[op13] M. J. Stokmaier, A. G. Class, and T. Schulenberg. "A hard optimisation test function with symbolic solution visualisation for fast interpretation by the human eye". In: *2013 IEEE Congress on Evolutionary Computation (CEC)*. 2013, pp. 2251–2258.

[op14] M. J. Stokmaier, A. G. Class, T. Schulenberg, and R. T. Lahey Jr. "Optimising acoustic resonators for sonofusion experiments with evolutionary algorithms". In: *Proceedings of the International Conference on the Applications of Computer Science and Mathematics in Architecture and Civil Engineering*. Bauhaus-University Weimar, July 2015.