TOWARDS RAPID LANGUAGE PORTABILITY OF SPEECH PROCESSING SYSTEMS

Tanja Schultz

Interactive Systems Laboratories, Carnegie Mellon University
E-mail: tanja@cs.cmu.edu

ABSTRACT

In recent years, more and more speech processing products in several languages have been widely distributed all over the world. This fact reflects the general believe that speech technologies have a huge potential to let everyone participate in today's information revolution and to bridge the language barriers. However, the development of speech processing systems still requires significant skills and resources to be carried out. With some 4500- 6000 languages in the world, the current cost and effort in building speech support is prohibitive to all but the top, most economically viable languages. In order to overcome these limitations, our research centers around the development of new algorithms and tools to rapidly port speech processing systems to new languages. This paper focuses on our approaches to create acoustic models, pronunciation dictionaries, and language models in new languages with only limited or no data resources available in the language of question. For this purpose we developed language independent and language adaptive acoustic models, investigated pronunciation dictionaries which can be directly derived from the written form and propose cross-lingual language model adaptation. The approaches are evaluated on our multilingual text and speech database *GlobalPhone* which covers more than 15 languages of the world.

1 INTRODUCTION

The global trend to small, mobile devices in conjunction with today's computerization is one of the major driving force in speech and language processing since speech is the most natural front-end to communicate with and through computers. To date speech-driven applications have only been built in the most economically viable languages, however we believe that speech-driven applications will only be successful, if they are provided in the user's native tongue. Therefore, speech processing is required to become available in a huge number of languages and even spoken dialects in order to reach the majority of people. This includes languages in which only few or no resource are available. As a consequence, a massive reduction of effort in terms of time and costs is necessary to speed up the development of recognizers in new tasks and languages.

Our fundamental research goal is to reveal techniques and algorithms that allow to rapidly develop automatic speech processing systems in many languages. We successfully built speech and text data resources in a large variety of languages that serves as one basis of our research. Within this framework we successfully developed language independent acoustic models to rapidly bootstrap acoustic models in new languages. We furthermore developed a fully automatic generation scheme for pronunciation dictionaries, and recently started to investigate crosslingual languages model adaptation. Within the recently awarded NSF project SPICE (**S**peech **P**rocessing: **I**nteractive **C**reation and **E**valuation toolkit), we will tackle one of the major obstacles for the development of speech processing components in a new language, i.e. the lack of human language technology experts. We will overcome this bottleneck by breaking the link between language and technology expertise. This will be implemented by providing innovative methods and tools for unskilled users to develop speech processing models, collect appropriate data to build these models, and evaluate the results allowing iterative improvements. The evaluation is planned to be performed with a strong focus on Indian languages.

2 THE GLOBALPHONE PROJECT

The increasing demand for rapid deployment of speech processing systems in new languages is accompanied by the need for a multilingual speech and text database that covers a broad variety of languages while being uniform across languages. Uniformity here refers to the total amount of text and audio per language as well as to the quality of data, such as recording conditions (noise, channel, microphone etc.), collection scenario (task, setup, speaking style etc.), and transcription conventions. Only uniform data allow the development of global phone sets and enable the comparison of speech and/or text across languages. To train and evaluate large vocabulary continuous speech recognition systems, dozens of hours of audio data from many speakers together with transcripts are required for acoustic modeling, and text data of millions of written words need to be available for language modeling. Furthermore, research in multilingual speech processing requires databases that cover the most relevant languages.

This section briefly describes the design, collection, and current status of the multilingual database *GlobalPhone*, a speech and text database available in 15 languages: Arabic, Chinese (Mandarin and Shanghai), Croatian, Czech, French, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. In total, the corpus contains more than 300 hours of transcribed speech spoken by more than 1500 native, adult speakers and will soon be available from ELRA [ELRA].

Language	*Number Speakers*	*Audio [hours]*	*Spoken Words*
Arabic	170	35	i.p.
Ch-Mandarin	132	31	263k
Ch-Shanhai	41	10	95k
Croatian	92	16	120k
Czech	102	29	220k
French	94	25	250k
German	77	18	151k
Japanese	144	34	268k
Korean	100	21	117k
Portuguese	101	26	208k
Russian	106	22	170k
Spanish	100	22	172k
Swedish	98	22	184k
Tamil	49	i.p.	i.p.
Turkish	100	17	113k
Total	1506	328	2331k

Table 1: The GlobalPhone corpus (i.p. = in progress)

GlobalPhone is designed to provide read speech data for the development and evaluation of large continuous speech recognition systems in the most widespread languages of the world, and to provide a uniform, multilingual speech and text database for language independent and language adaptive speech recognition as well as for language identification tasks. The entire GlobalPhone corpus enables the acquisition of acoustic-phonetic knowledge of 15 languages. The languages were selected considering criteria such as: (1) Size of speaker population, (2) Political and economic relevance, (3) Geographic coverage, (4) Phonetic coverage, (5) Orthographic script variety, and (6) Morphologic variety. However, size of speaker population and language relevance was favored above geographic coverage. Some languages were collected to study cross-language portability within language families. Considering the fact that English is already available in a very similar framework (Wall Street Journal), the database covers 9 out of the 12 most frequent languages of the world. In each language about 100 sentences were read from each of 100 speakers. This corresponds to 20 hours spoken speech, i.e. around 10,000 utterances or roughly 100,000 spoken words per language. The read texts were selected from national newspapers available via Internet to provide a large vocabulary (up to 65,000 words). The read articles cover national and international political news as well as economic news from 1995-1998. The chosen domain allows for additional collection of suitable large text corpora for language modeling by web crawling. The speech is available in 16bit, 16kHz mono quality, recorded with a close-speaking microphone (Sennheiser 440-6) in a quiet environment and same recording equipment for all languages. All GlobalPhone data were collected in the home countries of the native speakers to avoid artifacts which might occur when living in a non-native environment. The transcriptions are internally validated and supplemented by special markers for spontaneous effects like stuttering, false starts, and non-verbal effects like laughing and hesitations. The transcripts are available in the original orthographic script, but were additionally mapped into a romanized form. Speaker information like age, gender, occupation, etc. as well as information about the recording setup complement the database.

Table 1 shows the current status of the GlobalPhone corpus. The average length per turn is about 9sec. The average number of words spoken in a turn is about 19 units, but varies across languages with the length of the word unit (segmentation). For more details about the database please refer to [Schultz2002].

3 LANGUAGE INDEPENDENT ACOUSTIC MODELING

Global Phoneme Inventory

Our research in design and implementation of a language independent or *global phoneme set* is based on the assumption that the articulatory representations of phonemes are so similar across languages, that phonemes can be considered as units which are independent from the underlying language. As a consequence we unify the language specific phoneme inventories of languages into one global set. This idea is a fundamental aspect of the International Phonetic Association [IPA1993] and has been embodied in the research of language identification by [Andersen1997] and [Corredor-Ardoy1997].

In [Schultz2001] we defined a global unit set for 12 languages (Chinese, English, French, German, Japanese, Korean, Croatian, Portuguese, Russian, Spanish, Swedish, and Turkish) based on the IPA scheme and developed acoustic models for speech recognition. Sounds of different languages, which are represented by the same IPA symbol, share one common unit, so-called *IPA-uni*t, in this global unit set. According to this idea we differentiate between the group of language

independent *poly-phonemes* containing phonemes occurring in more than one language, and remaining groups of language dependent *mono-phonemes*. Table 2 summarizes the poly-phonemes and mono-phonemes which cover 9 of the 12 most widespread languages in the world. For each poly-phoneme the upper half of Table 2 reports the number of languages which share one phoneme. The lower half of Table 2 contains the number and type of mono-phonemes for each language. In total, the global unit set consists of 485 language dependent phonemes which had been shared into 162 classes. Therefore, on average, each phoneme of our global unit set is shared by 3 languages. We found that this phoneme *share factor* increases with the number of languages, and also strongly depends on the involved languages, implying that the phoneme inventories of some languages are quite similar while others are not [Schultz2001]. The global unit set in conjunction with the acoustic models covering 12 languages of the world provides us with the optimal basis to select phonemes for new languages and use the corresponding language independent acoustic models as seeds for the acoustic models of the new language.

Shared by	#	Modeled Phonemes (IPA symbols)	
	83	Polyphonemes shared across ≥ 2 languages	
		Consonants	Vowels
All	4	m,n,s,l	-
11	7	p,b,t,d,k,g,f	-
10	3	-	i,u,e
9	6	ŋ,v,z,j	a,o
8	1	ʃ	-
7	3	r,h,tʃ	-
6	1	-	ɛ
5	9	ɲ,ʒ,x,ts,dʒ	i:,y,ø,ɔ
4	4	-	ɬ,ø,ɑ,ei
3	11	ʎ,w,ç	ɪ,u:,e:,œ,o:,æ,ai,au
2	34	p^h,t^h,d^j,k^h,g^j,ɪr,ɾ$_r$, θ,ð,s^j,z^j,ʂ,ɻ,ts^h,$tʃ^h$	ˈi,y:,ɯ,ʊ,ˈe,œ:,ø:,a:,ˈa,ɑ:, ˈu,ˈo,ai,au,ia,io,eu,oi,ou
	79	Monophonemes belonging to *one* language	
		Consonants	Vowels
CH	15	$ʈ_ʂ$,$ʈ^h_ʂ$,cɕ,$cɕ^h$	iʊ,iɛ,ua,uɛ,uɔ,ya,yɛ, iao,uei,uai,iou
EN	5	$ɾ_d$	ʌ,ɔ˞,ɔi,ə˞
FR	5	ɳ	ɛ̃,œ̃,ɑ̃,ɔ̃
GE	3	-	ɐ,ʏ,ɔʏ
JA	2	ʔ	ɯ:
KO	14	$p^=$,p',$t^=$,t',$k^=$,k', s',$ɕ'^h$	ie,iɘ,iu,ɨi,oa,uɘ
KR	1	$dʒ^j$	-
PO	8	-	ĩ,ũ,ɛ̃,ɔ̃,ɐ̃,ew,ow,aw
RU	15	p^j,b^j,f^j,m^j,r^j,v^j, $ʃ^j$,$ʒ^j$,$ɻ^j$,ʃʃ,ʃʃ^j	ja,jɛ,jo,ju
SP	2	β,ɣ	-
SW	9	ʈ,ɖ,ɳ,ɭ,ks	œ:,æ:,ʉ:,ɵ
TU	0	-	-
Σ	162	Silence and noises shared across languages	

Table 2: Global Phoneme Inventory

Rapid Adaptation of Acoustic Models

Based on the described global unit set together with created monolingual systems we investigate different methods to combine the acoustic models of varied languages to one multilingual acoustic model. The main goals of the model combination were the reduction of the overall amount of acoustic model parameters and the improvement of the model robustness for language adaptation purposes. We applied the language independent acoustic models to initialize the acoustic models of the target language recognizer using seed models developed for other languages [Schultz2001]. Previous approaches for language adaptation have been limited to context independent acoustic models. Since for the language dependent case wider contexts increase recognition performance significantly, we investigate whether such improvements extend to the multilingual setting. The use of wider context windows raises the problem of phonetic context mismatch between source and target languages. To measure this mismatch we define the coverage coefficient. In order to approach the mismatch problem we introduce a method for polyphone decision tree adaptation where the clustered multilingual polyphone decision tree is adapted to the target language by restarting the decision tree growing process according to the limited adaptation data available in the target language [Schultz2000].

We investigated the benefit of the acoustic model combination and the polyphone decision tree specialization (PDTS) for the purpose of adaptation to the Portuguese language. Figure 1 summarizes the experiments which have been performed to improve the Portuguese LVCSR system. The row labeled *SystemId* gives the name which is used to identify the developed systems. The row *Data* refers to the amount of adaptation data (0-90 minutes of spoken speech). *Quality* explains whether the phonetic alignments are initially created based on the multilingual recognition engine or assumed to be available in good quality. The term *Method* is related to the porting approach which is applied: Cross-language transfer (CL), adaptation (Viterbi or MLLR), and bootstrapping technique (Boot). *Viterbi* refers to one iteration of Viterbi training along the given alignments. *MLLR* is the Maximum Likelihood Linear Regression [Leggetter1995], and *Boot* refers to the iterative procedure: creating alignments, Viterbi training, model clustering, training, and writing improved alignments. The item *Tree* describes the origin of the polyphone decision tree: '–' refers to context independent modeling, *LI* is the generic language independent polyphone decision tree of a mixed acoustic model system, *LD* is the language dependent tree which

is built exclusively on Portuguese data, and *PDTS* refers to the adapted LI polyphone tree after applying PDTS.

In summary, we achieved 19.6% word error rate when adapting language independent acoustic models to the Portuguese language using only 90 minutes of spoken Portuguese speech. This compares to 19.0% of a full trained system on 16.5 hours of spoken Portuguese speech. The adaptation procedures runs on a 300MHz SUN Ultra and takes only 3-5 hours real-time. As a consequence the introduced techniques allow to set up LVCSR systems in a new target language without the need of large speech databases in that language. In combination with an automatic generation of pronunciation dictionaries (see section 4) and a method to generate a language model for example by fully automatically downloading appropriate text resources from the web (see section 5), a speech recognition could be developed very efficiently.

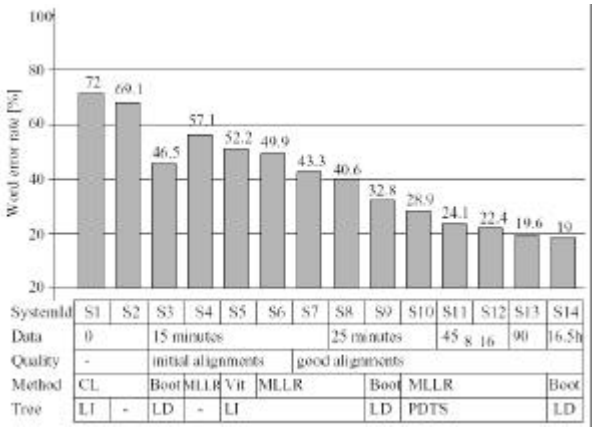

Figure 1: Language Adaptation to Portuguese

4 AUTOMATIC GENERATION OF PRONUNCIATION DICTIONARIES

Besides acoustic modeling, the pronunciation dictionary is another core component of a speech recognition system. Its purpose is to map the written form of vocabulary entries to units which model their actual acoustic realization. Usually, phonemes or sub-phonetic units are used as acoustic model units. The performance of a speech recognizer heavily depends on the quality of the pronunciation dictionary and best results are usually achieved with hand-crafted dictionaries. However, this manual approach is very time and cost consuming especially for large vocabulary speech recognition. Moreover, as applications become interactive, the demand for on-the-fly dictionary expansion increases, as for example in voice driven cell phone applications which support name dialing.

Consequently, methods to automatically create dictionaries are necessary in all those cases where no language expert knowledge is available or time and cost limitations prohibit the manual creation. Several methods have been introduced in the past, especially in the context of text-to-speech processing. Here, methods are mostly based on finding rules for converting the written form of a word into its phonetic transcription, by either applying rules as for example in [Black1998] or by statistical approaches [Besling1994]. In speech recognition only very few approaches have been investigated so far [Singh2002] but recently, the use of graphemes as modeling units for speech recognition has been proposed [Kanthak2002].

The idea of using graphemes as model units, i.e. speech recognition based on the orthography of a word, is very appealing especially in the context of rapid portability to new languages since it makes the generation of a pronunciation dictionary a very straightforward task. However, it requires that (1) the orthographic representation of a word is given and (2) the relation between the written and the spoken form is reasonably close. Today some hundred different writing systems exist in the world and the majority are phonological scripts [Weingarten2003], i.e. they link the letters with the sounds. Phonological scripts are divided into syllable based scripts (e.g. Japanese kana) and alphabet scripts. Most alphabets consist of 20-30 symbols ranging from 11 (Rotokas alphabet) to 74 symbols (Khmer alphabet). The most widely used script is probably the roman script which was taken over from the Etruscan. Due to its widespread use, languages without written forms are likely to adopt some variation of the roman script (as happened for example in Mapudungun). As a consequence it is reasonable to assume that we can reach a very large number of languages with the grapheme based approach. Furthermore, we will show in the next section that the grapheme-based approach is not only feasible for languages with roman script but also for other scripts such as Cyrillic and Thai.

Grapheme-based speech recognition

The performance of a grapheme based speech recognizer is highly influenced by the closeness of the grapheme-to-phoneme relation. This relationship varies widely across languages. Some languages such as Spanish and Finnish have an almost perfect one-to-one relation, while others such as English show major irregularities. The reasons for irregularities are manifold, mostly since the script is not appropriate for a particular language or did not follow the modifications of the spoken language. In only few cases the alphabet had been re-adapted (e.g. Turkish) or invented (e.g. Korean) to better represent the spoken form.

We investigated the potential of the grapheme based modeling approach in the context of rapid portability to new languages. For this purpose we selected a variety of languages from our GlobalPhone corpus: English, German, and Spanish, as examples of the roman script where English shows the weakest grapheme-to-phoneme correspondence, Spanish shows the strongest, and German lies somewhere in between. Additionally, we investigated the potential of this approach on languages written with other than roman scripts, namely Russian and Thai.

The first and the second column in Table 3 compares the performance of phoneme based with grapheme based speech recognizers for these five languages. All settings and components of the speech engine are the same except for the acoustic model and dictionary. Also the parameter size is the same. The results show that grapheme based systems perform significant worse for languages with poor grapheme-to-phoneme relation such as English, but achieve comparable results for closer relations such as Spanish and Russian. In case of German we even see a gain by using graphemes over phonemes which is most likely due to the more consistent dictionary. For more details on our studies please refer to [Mimer2004] for English and German, [Killer2003] for Spanish, and [Stüker2004] for Russian. The results for Thai are preliminary and we expect to significantly reduce the gap between the phoneme and the grapheme based approach in the near future.

The absolute performance differences across the languages are due to a variety of factors such as systems' maturity, different out-of-vocabulary rates due to morphology and/or vocabulary size, and language model training corpus size, to name only a few.

Language	*Phonemes*	*Graphemes*	*Tree-Tied Gr*
English	11.5	19.5	18.6
German	15.6	14.0	12.7
Spanish	24.5	26.8	-
Russian	33.0	36.4	32.8
Thai	14.0	26.4	-

Table 3: Phoneme vs Grapheme based ASR [WER in %]

Tree-Tied Graphemes

Recent results in pronunciation modeling seem to indicate that pronunciation variants should not be explicitly modeled through phoneme string variations but rather implicitly by the use of single pronunciation dictionaries [Hain2002] and parameter sharing across phonetic models [Saraclar2000]. In this sense, a grapheme based dictionary is a single pronunciation dictionary in its purest form.

Traditionally the acoustic units are modeled using polyphones i.e. phonemes in the context of neighboring phonemes. Since the number of polyphones even for a very small context width is too large to allow a robust model parameter estimation, context dependent models are usually clustered into classes using a decision tree based state tying [Young1994]. Due to computational and memory constraints, those cluster trees are grown for each phoneme sub-state. However, this scheme prohibits parameter sharing across polyphones of different center phonemes. This constraint is lifted by the enhanced tree clustering as described in [Yu2003]. In this scheme a single decision tree is constructed for all sub-states of all phonemes and thus allows a flexible sharing across phonemes. We applied this clustering scheme to grapheme based speech recognition. Here a dictionary can not capture the fact that (a) the same grapheme might be pronounced in different ways depending on the context and (b) that different graphemes might be pronounced the same way depending on the context. The traditional clustering procedure is able to deal with the effects of (a) but in order to handle the implications of (b) and make the best use of the available training data at the same time, the enhanced tree clustering is needed. We applied the enhanced tree tying to the languages German, English, and Russian. The results are presented in the third column of Table 3. They show that enhanced tree tying outperforms the standard decision tree clustering and thus indicate that sharing across graphemes captures the fact that different graphemes are pronounced similar depending on their context. With the enhanced tree clustering the grapheme based speech recognition outperforms the phoneme based approach in case of German and Russian, and closes the gap for English.

Additionally, we build language independent grapheme models by resembling our work on language independent phoneme acoustic models and investigated the potential for rapid adaptation to new languages [Killer2003]. The results show limited success confirming our suspicion that grapheme systems are rather consistent within a language but not across languages.

5 LANGUAGE MODELING

The main concern of (statistical) language modeling is to reliably estimate the probabilities of word sequences in the context of a particular language and/or domain. Many approaches had been proposed to tailor language models towards particular domains such as language model adaptation by text selection, or various interpolation schemes. Some methods have been introduced to transfer knowledge across languages such as the exploitation of parallel texts to project morphological

analyzers or POS-taggers [Yarowsky2001]. However, in those cases it is assumed that a large number of (bilingual) text data is available or has been collected for the language in question. In this section we outline ideas for language model creation in languages where only few data resources are available or time and cost limitations require a rapid deployment.

One promising approach is a crosslingual language model adaptation as proposed by [Kim2003]. The algorithm first identifies text data in a resource-rich language which are similar to the target language, then extracts useful statistics from those text data, and projects the statistics back into the target language. This approach uses Information Retrieval methods to find contemporaneous articles of source and target languages, derives a corpus aligned set of corresponding articles, and uses text translation to find semantically related translation pairs. Figure 2 shows the procedure with source language L1 and target language L2.

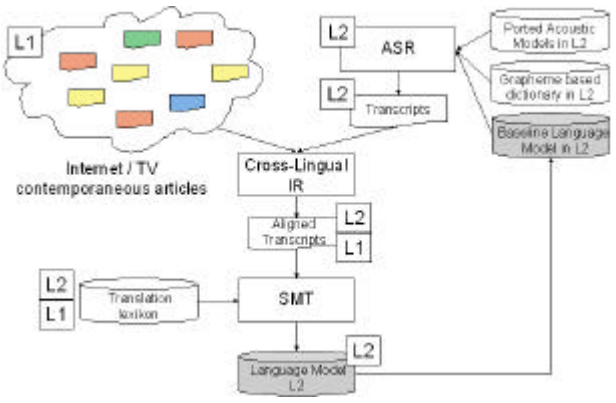

Figure 2: Crosslingual Language Model Generation

Another approach which is applicable for small domains is the usage of grammar based recognizers. Our results with multilingual language modeling for multilingual speech interfaces [Fügen2003] indicate that some text-based knowledge might be sharable across languages such as named entities. Using multilingual grammars would therefore be one way to transfer knowledge across languages. Grammars and statistic language models could also be intertwined to rapidly bootstrap larger domains from knowledge on smaller domains. We currently explore the described schemes to investigate their potential for rapid language model generation.

6 TOOLS FOR RAPID DEPLOYMENT

Speech recognition as well as speech synthesis have significantly improved over recent years in building recognizers and voices in new languages. However, in spite of comprehensive toolkits (e.g. Janus [Finke1997, Soltau2001] and Festvox [Festival1998, Festvox2000]), it is still a skilled job requiring significant effort from trained individuals. Deciding on a phone set, constructing a pronunciation lexicon, and designing a database that covers variation in languages, still requires more effort than many are willing or able to devote. The primary focus of SPICE (**S**peech **P**rocessing - **I**nteractive **C**reation and **E**valuation Toolkit for new Languages), a three years program sponsored by NSF, is to overcome this limitation by providing innovative methods and tools for naive users to develop speech processing models, collect appropriate data to build these models, and evaluate the results allowing iterative improvements [Spice]. Building on the existing GlobalPhone and FestVox projects, knowledge and data will be shared between recognition and synthesis such as phoneme sets, pronunciation dictionaries, acoustic models, and text resources. User studies will indicate how well speech systems can be build, how well tools support the efforts and what must be improved to create even better systems. This research increases the knowledge of how to rapidly create speech recognizers and synthesizers in new languages. Furthermore, archiving the data gathered on-the-fly from many native cooperative users will significantly increase the repository of languages and resources. We hope to revolutionize the speech system generation by integrating speech recognition and synthesis technologies into an interactive language creation and evaluation toolkit usable by unskilled users. Data and components for new languages will become available at large to let everybody participate in the information revolution, improve the mutual understanding, bridging language barriers, and thus foster the educational and cultural exchange.

7 CONCLUSIONS

We introduced techniques that allow to set up large vocabulary continuous speech recognition systems in a new target language without the need of large speech and text databases in that language in question. Our implementation of language independent acoustic models in combination with a grapheme based automatic dictionary generation shows very good results without the need of large language resources and language experts. We furthermore outlined ideas towards crosslingual language model adaptation making use of contemporaneous text articles from the internet and/or multilingual grammars. Based on the introduced technologies together with the implementation of interaction speech processing creation and evaluation toolkits we will soon be able to rapidly deploy speech processing systems without the need of language technology experts and without the need of large text and speech data and thus allow people from all different language background to participate in today's information revolution.

REFERENCES

[Andersen1997] O. Andersen, and P. Dalsgaard, *Language Identification based on Cross-language Acoustic Models and Optimised Information Combination*. Eurospeech, Rhodes 1997, pp. 67-70.

[Besling1994] S. Besling, *Heuristical and statistical Methods for Grapheme-to-Phoneme Conversion*, Konvens, Wien, Austria, p.23-31, 1994.

[Black1998] A. Black, K. Lenzo, and V. Pagel, *Issues in building general letter to sound rules*, Proceedings of the ESCA Workshop on Speech Synthesis, Australia., pp. 77–80, 1998.

[Corredor-Ardoy1997] C.Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, and L. Lamel, *Language Identification with Language-independent Acoustic Models*. Eurospeech, pp. 355-358, Rhodes, Greece, 1997.

[ELRA] European Language Resources Association (ELRA): http://www.icp.grenet.fr/ELRA/home.html

[Festival1998] A. Black, P. Taylor, and R. Caley, *The Festival Speech Synthesis System.* http://festvox.org/festival, 1998.

[Festvox2000] A. Black, K. Lenzo, *Building Voices in the Festival Speech Synthesis System.* http://festvox.org/bsv/, 2000.

[Finke1997] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, *The Karlsruhe-Verbmobil Speech Recognition Engine*, ICASSP, pp. 83–86, Munich, Germany, 1997.

[Fügen2003] C. Fügen, S. Stüker, H. Soltau, F. Metze, and T. Schultz, *Efficient Handling of Multilingual Language Models*. ASRU, St. Thomas, VI, 2003.

[Hain2002] T. Hain, *Implicit pronunciation modelling in ASR*, ISCA Pronunciation Modeling Workshop, 2002.

[IPA1993] IPA: *The International Phonetic Association (revised to 1993) - IPA Chart*, Journal of the International Phonetic Association 23, 1993.

[Kanthak2002] S. Kanthak and H. Ney, *Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition.* ICASSP, pp. 845-848, Orlando FL, 2002.

[Killer2003] M. Killer, S. Stüker, and Tanja Schultz. *Grapheme based Speech Recognition.* Eurospeech, Geneva, Switzerland, September 2003.

[Kim2003] W. Kim and S. Khudanpur, *Language Model Adaptation Using Cross-Lingual Information.* Eurospeech, 3129–3132, Geneva, Switzerland, 2003.

[Mimer2004] B Mimer, S. Stüker, and T. Schultz, *Flexible Tree Clustering for Grapheme-based Speech Recognition.* Elektronische Sprachverarbeitung (ESSV), Cottbus, Germany, September 2004.

[Saraclar2000] M. Saraclar, H.J. Nock, and S. Khudanpur, *Pronunciation Modeling By Sharing Gaussian Densities Across Phonetic Models*, Computer Speech and Language, vol. 14, pp. 137-160, 2000.

[Schultz2000] T. Schultz and A. Waibel: *Polyphone Decision Tree Specialization for Language Adaptatio*n. ICASSP, Istanbul, Turkey, June 2000.

[Schultz2001] T. Schultz and A. Waibel, *Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition*, Speech Communication, Volume 35, Issue 1-2, pp 31-51, August 2001.

[Schultz2002] T. Schultz, *GlobalPhone: a Multilingual Speech and Text Database developed at Karlsruhe University.* ICSLP, Denver, CO, September 2002.

[Singh2002] R. Singh, B. Raj and R. M. Stern, *Automatic Generation of Subword Units for Speech Recognition Systems*, IEEE Transactions on Speech and Audio Processing, Vol. 10, p. 98-99, 2002.

[Soltau2001] H. Soltau, F. Metze, C. Fügen, and A. Waibel, *A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment*, Proceedings of the ASRU, Madonna di Campiglio Trento, Italy, December 2001.

[Spice] http://www.is.cs.cmu.edu/Spice

[Stüker2004] S. Stüker and T. Schultz, *A Grapheme Based Speech Recognition System for Russian*, Specom 2004, St. Petersburg, Russia, September 2004.

[Weingarten2003] R. Weingarten, http://www.ruediger-weingarten.de/Texte/Latinisierung.pdf, University of Osnabrück, 2003.

[Young1994] S. Young, J. Odell, and P. Woodland, *Tree-based state tying for high accuracy acoustic modelling*, Proceedings of the ARPA HLT Workshop, Princeton, New Jersey, March 1994.

[Yu2003] H. Yu and T. Schultz, *Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversational Speech Recognition*, Eurospeech, Geneva, Switzerland, September 2003.