Multilingual Speech Processing in the context of Under-resourced Languages

Tanja Schultz
Karlsruhe University and Carnegie Mellon University

Tutorial at SLTU '08, Monday, May 5th 2008, Hanoi

Carnegie Mellon

Karlsruhe Institute of Technology

Tutorial Outline

- Introduction and Motivation
  - Motivation
  - History and Leveraged Work
  - Rapid Language Adaptation Server: Spice
- SPICE in detail
  - Text collection & Prompt Selection
  - Phone set specification, Lexical construction
  - ASR Bootstrap & training
  - Language model, TTS Voice building
  - Testing and Tuning
- Latest Experiments and Results
  - Lessons Learnt from past studies
- Conclusions & Next Steps

Introduction & Motivation

- Many Languages – so what?
  - Growing Language Diversity on the web
  - Why do we need Speech Processing in many languages?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Language Extinction
  - Do we care? What can we do about?
- Challenges of Multilingual Speech Processing
  - Lack of Resources
  - Lack of Experts
- Solutions
  - Prior Work: GlobalPhone and FestVox
  - Intelligent Learning Systems
  - Rapid Language Adaptation Server

Introduction & Motivation

- Many Languages – so what?
  - Growing Language Diversity on the web
  - Why do we need Speech Processing in many languages?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Language Extinction
  - Do we care? What can we do about?
- Challenges of Multilingual Speech Processing
  - Lack of Resources
  - Lack of Experts
- Solutions
  - Prior Work: GlobalPhone and FestVox
  - Intelligent Learning Systems
  - Rapid Language Adaptation Server

Many Languages – So What?

Do we really need Speech Processing in many languages?

Myth: "Everyone speaks English, why bother?"

- ➔ NO: About 6900 different languages in the world
- ➔ Increasing number of languages on the web
- ➔ Humanitarian and military needs
  - ➔ Rural areas, uneducated people, illiteracy

Why is this an research issue?

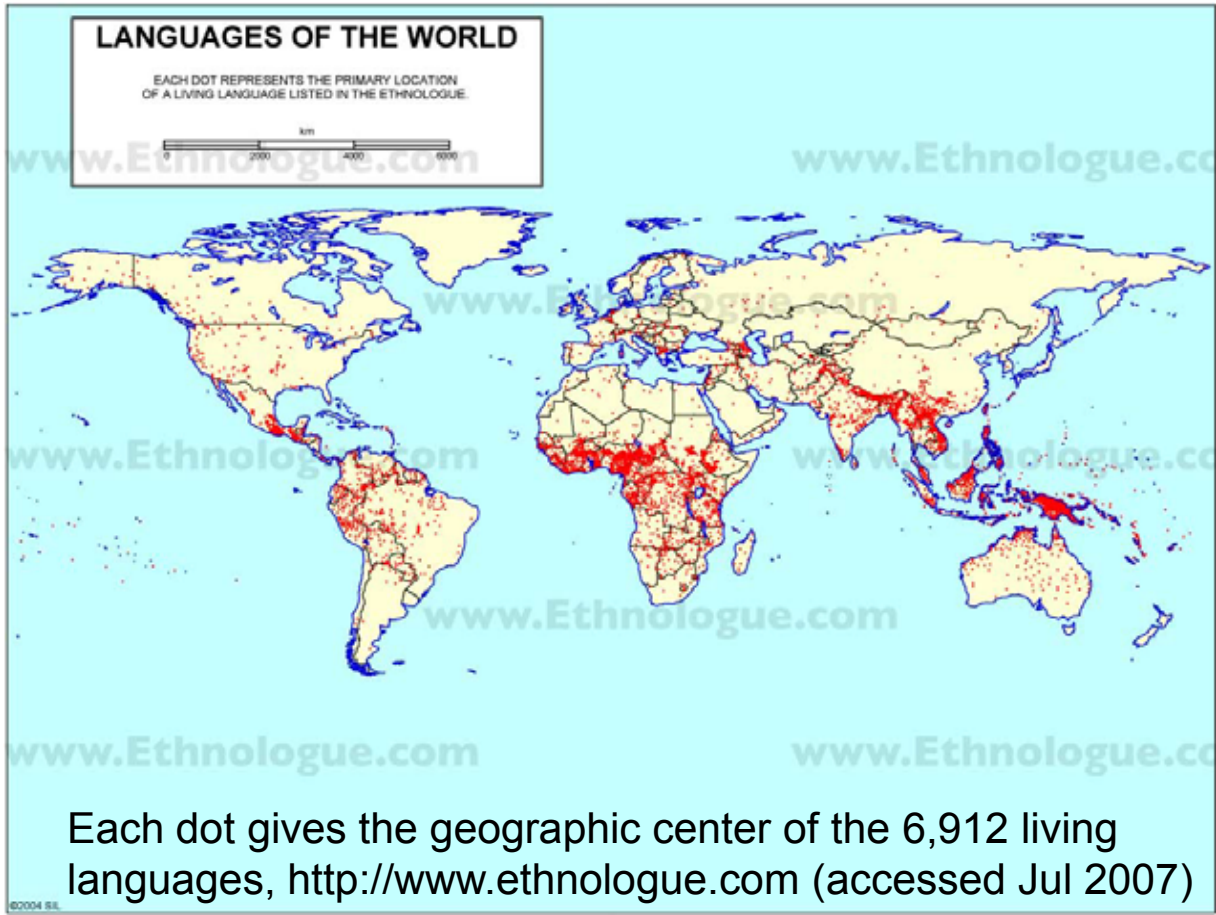
Myth: "It's just retraining on foreign data – simple!"

- ➔ NO: Other languages bring unseen challenges, for example:
  - ➔ different scripts, no vowelization, no writing system
  - ➔ no word segmentation, rich morphology,
  - ➔ tonality, click sounds,
  - ➔ social factors: trust, access, exposure, cultural background

Everyone speaks English, why bother?

- Huge number of Languages in the world: **6912**
- Language is not only a communication tool but fundamental to cultural identity and empowerment

- Treat linguistic diversity as we treat bio-diversity (David Crystal)
- The strongest eco systems are the most diverse
- Cultures, ideas, memories are transmitted ***through language***

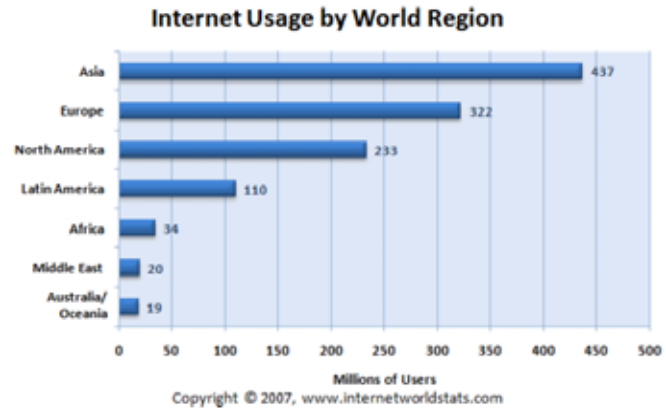


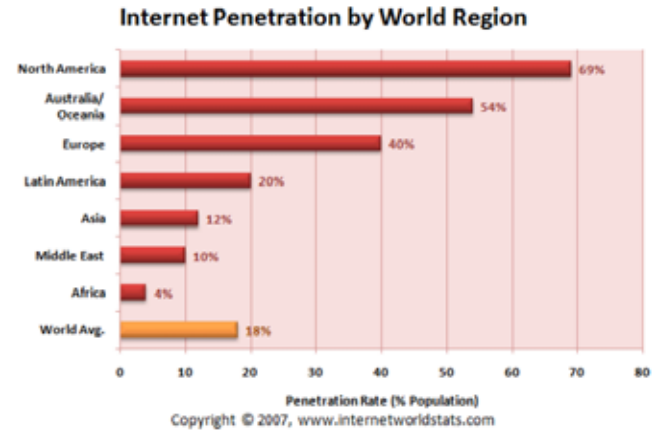
Each dot gives the geographic center of the 6,912 living languages, http://www.ethnologue.com (accessed Jul 2007)

Growing Diversity, no Uniformity!

Diversity of Languages in the Internet grows rapidly

- Top-10: 200%
- All others: 440%
- Chinese: 414%
- Arabic: 940%





Top Ten Languages Used in the Web
(Number of Internet Users by Language)

TOP TEN LANGUAGES IN THE INTERNET	% of all Internet Users	Internet Users by Language	Internet Penetration by Language	Internet Growth for Language (2000 - 2007)	2007 Estimate World Population for the Language
English	31.7 %	365,893,996	17.9 %	157.7 %	1,042,963,129
Chinese	31.7 %	166,001,513	12.3 %	413.9 %	1,351,737,925
Spanish	8.8 %	101,539,204	22.9 %	311.4 %	442,525,601
Japanese	7.5 %	86,300,000	67.1 %	83.3 %	128,646,345
German	5.1 %	58,981,592	61.1 %	112.9 %	96,488,326
French	5.1 %	58,456,702	15.1 %	379.2 %	387,820,873
Portuguese	4.1 %	47,326,760	20.2 %	524.7 %	234,099,347
Korean	3.0 %	34,120,000	45.6 %	79.2 %	74,811,368
Italian	2.7 %	31,481,928	52.9 %	138.5 %	59,546,696
Arabic	2.5 %	28,782,300	8.5 %	940.5 %	340,548,157
TOP TEN LANGUAGES	84.8 %	978,883,995	19.0 %	198.0 %	5,159,187,766
Rest of World Languages	15.2 %	175,474,783	12.4 %	440.3 %	1,415,478,651
WORLD TOTAL	100.0 %	1,154,358,778	17.6 %	219.8 %	6,574,666,417

(*) NOTES: (1) Internet Top Ten Languages Usage Stats were updated for June 30, 2007. (2) Internet Penetration is the ratio between the sum of Internet users speaking a language and the total population estimate that speaks that specific language. (3) The most recent Internet usage information comes from data published by Nielsen//NetRatings, International Telecommunications Union, Computer Industry Almanac, and other reliable sources. (4) World population information comes from the world gazetteer web site. (5) For definitions and navigation help, see the Site Surfing Guide. (6) Stats may be cited, stating the source and establishing an active link back to Internet World Stats.

Why Speech Processing?

So we need language support but why *Speech*?

- Computerization: Speech is *the* key technology
  - Ubiquitous Information Access: on the go, phone-based
  - Mobile Devices: Too small and cumbersome for keyboards

- Globalization:
  - Cross-cultural Human-Human Interaction
    - Multilingual Communities: EU, South Africa, …
    - Humanitarian needs, disaster, health care
    - Military ops, communicate with local people
  - Human-Machine Interfaces
    - People expect speech-driven applications in their mother tongue

⇒ **Speech Processing in multiple Languages**

ML Speech Processing – Research Issue?

It's just retraining on foreign data - no science!

No New language – new challenges

- Writing system: different or no script, no vowelization, G-2-P
- Word segmentation, morphology
- Sound system: tonals, clicks

No Different Cultures – social factors

- trust, access, exposure, background

No Lack of Data and Resources

- Audio recordings, corresponding transcripts
- Pronunciation Dictionaries, Lexicon
- Text corpora, parallel bilingual data

No Lack of Experts

- Technology experts without language expertise
- Native language experts without technology expertise

Introduction & Motivation

- Many Languages – so what?
  - Growing Language Diversity on the web
  - Why do we need Speech Processing in many languages?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Language Extinction
  - Do we care? What can we do about?
- Challenges of Multilingual Speech Processing
  - Lack of Resources
  - Lack of Experts
- Solutions
  - Prior Work: GlobalPhone and FestVox
  - Intelligent Learning Systems
  - Rapid Language Adaptation Server

Language Characteristics

➔ Prosody, Tonality: Stress, Pitch, Lenght pattern, Tonal contours
(e.g. Mandarin 4, Cantonese 8, Thai & Vietnamese 5)

➔ Sound system: simple vs very complex sound systems
(e.g. Hawaiian 5V+8C vs. German 17V+3D+22C)

➔ Phonotactics: simple syllable structure vs complex consonant clusters
(e.g. Japanese Mora-syllables vs. German pf,st,ks)

➔ Segmentation: Written form separate words by white space?
(NO: Chinese, Japanese, Thai, Vietnamese)

➔ Morphology: short units, compounds, agglutination

English: Natural segmentation into short units – great!

German: Compounds – not quite so good

Donau-dampf-schiffahrt**s**-gesellschaft**s**-kapitän**s**-mütze …

Turkish: Agglutination – looooong phrases

Osman-lı-laç-tır-ama-yabil-ecek-ler-imiz-den-mi**ş**-siniz

behaving as if you were of those whom we might consider not converting into Ottoman

Writing Systems

Writing systems – basic unit is a Grapheme:

Logographic: based on semantic units, grapheme represents meaning

Chinese: >10.000 *hanzi*; Japanese ~7000 *kanji,* Korean to some extend

Phonographic: based on sound units, grapheme represents sound

Segmental: grapheme roughly corresponds to phonemes

Latin (190), Cyrillic (65), Arabic (22) graphems

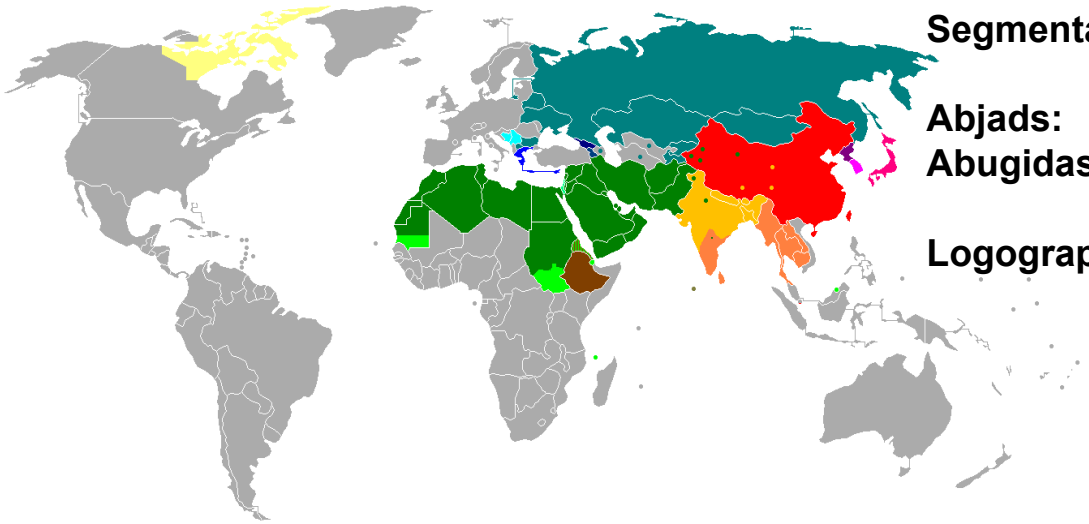
Abjads = consonantal segmental phonographic, e.g. Arabic

Syllabic: grapheme represents entire syllable, e.g. Japanese *kana*

Abugidas = mix of segmental and syllabic systems

Featural: elements smaller than phone, e.g. articulatory features

e.g. Korean: ~5600 *gulja*



Wikipedia: August 2007

Scripts – Examples

العربى български català 中国话 hrvatski česky
english ελληνικα עברית हिंदी italiano 日本語
한국어 românește русский српски ภาษาไทย

Scripts of some languages: Arabic, Bulgarian, Catalan, Chinese, Croatian, Czech, English, Greek, Hebrew, Hindi, Italian, Japanese, Korean, Romanian, Serbian, Thai

How many languages do have a written form?

- Omniglot lists about 780 languages that have scripts
- True number might be closer to 1000
  (Source Simon Ager, 2007, www.omniglot.com)

➔Logographic scripts, mostly 2 representatives:
  - Chinese: ~ 10.000 hanzi,
  - Japanese: ~7000 kanji (+ 3 other scripts ☺)

➔ Phonographic:
  - Korean: ~5600 gulja,
  - Arabic, Devanagari, Cyrillic, Roman: ~100 characters

Grapheme-to-Phoneme Relation

Grapheme-to-Phoneme (Letter-to-Sound) Relationship:

Logographic: NO relationship at all

concern for Chinese, Japanese, Korean

Phonographic: segmental: close – far – complicated

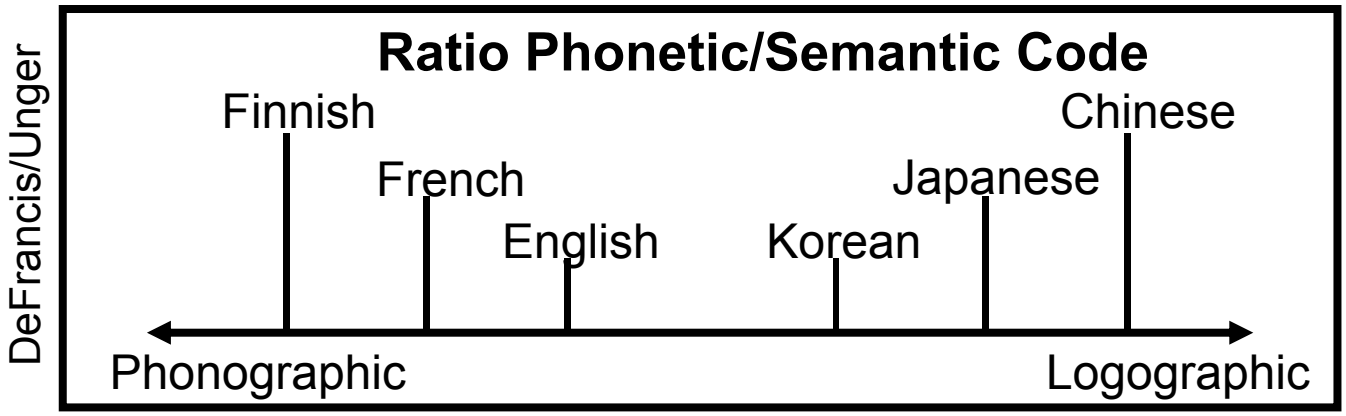
e.g. Finnish, Spanish: more or less 1:1, -- English: try „Phydough“

Phonographic: segmental – consonantal

e.g. Arabic: no short vowels written

Phonographic: syllabic

e.g. Thai, Devanagari: C-V flips



➔ Automatic Generation of Pronunciations might get complicated

Introduction & Motivation

- Many Languages – so what?
  - Growing Language Diversity on the web
  - Why do we need Speech Processing in many languages?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Language Extinction
  - Do we care? What can we do about?
- Challenges of Multilingual Speech Processing
  - Lack of Resources
  - Lack of Experts
- Solutions
  - Intelligent Learning Systems
  - Prior Work: GlobalPhone and FestVox
  - Rapid Language Adaptation Server

One more Reason for MLSP …

6900 Languages in the world …. BUT

- Extinction of languages on massive scale
  (David Crystal, Spotlight 3/2000)
- Half of all existing languages die out over next century
  ⇒ On Average: Every two weeks one language dies!
- Survey Feb 1999 from Summer Institute of Linguistics

51 languages with 1 speaker left

28 of those in Australia alone

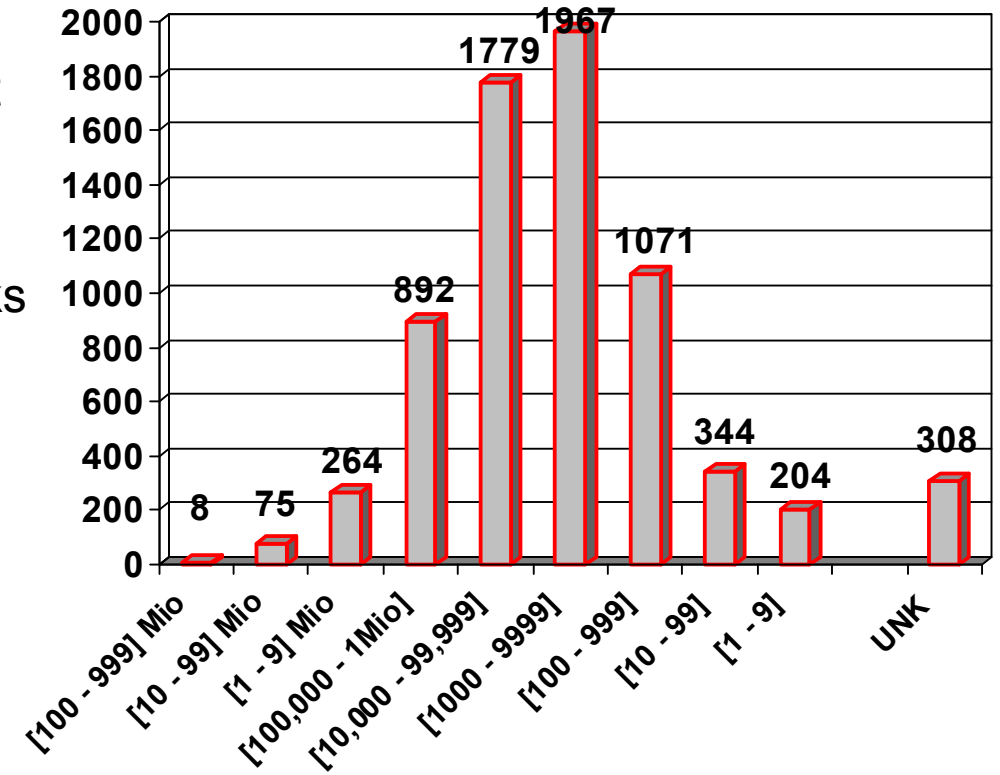
500 languages with < 500 spks

1500 languages with < 1000 spks

3000 languages with < 10.000

5000 languages with < 100.000

96% of world's languages are spoken by only 4% of its people



Carnegie Mellon

The Future of Language

Is a language with 100.000 speakers safe?

- Survival for generations depends on pressure imposed on language
- Dominance of another language, Attitude of the speakers
- Example Breton: beginning of 20th century has 1 Mio speakers, now down to 250.000; Without effort Breton could be gone in 50 years

Reasons that languages die:

- Disaster: Earthquake on Papua New Guinea: Sissano, Warapu, Arop
- Genocide: 90% America's natives died within 200 years Europeans
- Cultural assimilation: Colonialism, Suppression, Assimilation:
  - (1) Political, social, economic pressure to speak the dominant language,
  - (2) Emerging bilingualism,
  - (3) self-conscious semilingualism, (4) monolingualism

Why should we care?

- Massive death of languages reduces the diversity
- Bio-diversity has been accepted to be a good thing
- Maybe we should accepted this for language diversity (D. Crystal)

What can we do?

What do we learn from other languages?

- Intellectual issues: increase awareness of world history such as movements of early civilization
- Practical issues: medical practices, alternative treatment forms
- Literature … but also new things about the language itself
- Slovakian proverb: “with each newly learned language you acquire a new soul”

How to save endangered languages:

- Community itself must want it, Surrounding culture must respect it
- Funding for courses, materials, and teachers, support the community
- Get linguists into the field, publish information, grammars, dictionaries

Costs associated:

- Depends on conditions (written vs. unwritten languages, etc.)
- Crystal estimates about $80.000 / year per language
- 3000 endangered languages is about $700Mio …
- Organizations to raise funds
  - Foundation of endangered languages (FEL), UNESCO project

Introduction & Motivation

- Many Languages – so what?
  - Growing Language Diversity on the web
  - Why do we need Speech Processing in many languages?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Language Extinction
  - Do we care? What can we do about?
- Challenges of Multilingual Speech Processing
  - Lack of Resources
  - Lack of Experts
- Solutions
  - Intelligent Learning Systems
  - Prior Work: GlobalPhone and FestVox
  - Rapid Language Adaptation Server

Challenges of MLSP

- Lack of Resources: Stochastic approach needs **many** data
  - Hundreds of hours audio recordings and corresponding transcriptions
    Audio data ≤ 40 languages; Transcriptions take up to 40x real time
  - Pronunciation dictionaries for large vocabularies (>100.000 words)
    Large vocabulary pronunciation dictionaries ≤ 20 languages
  - Mono- and bilingual text corpora: few language pairs, pivot mostly English
- Algorithms are language independent – MLSP is not!
  - Other Languages bring unseen challenges (segmentation, G2P, etc.)
  - Have we already seen ALL or MOST of the language characteristics?
- Social and Cultural Aspects
  - Non-native speech and language, code switching
  - Combinatorical explosion (domain, speaking style, accent, dialect, ...)
  - Few native speakers at hand for minority (endangered) languages
  - Do we have the right data?
- Lack of Language Experts
  - Bridge the gap between technology experts and language experts

Introduction & Motivation

- Many Languages – so what?
  - Growing Language Diversity on the web
  - Why do we need Speech Processing in many languages?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Language Extinction
  - Do we care? What can we do about?
- Challenges of Multilingual Speech Processing
  - Lack of Resources
  - Lack of Experts
- Solutions
  - Intelligent Learning Systems
  - Prior Work: GlobalPhone and FestVox
  - Rapid Language Adaptation Server

One Solution: Learning Systems

⇒ Intelligent systems that learn a language from the user

- Efficient learning algorithms for speech processing
  - Learning:
    - Interactive learning with user in the loop
    - Statistical modeling approaches
  - Efficiency:
    - Reduce amount of data (save time and costs): at least by factor of 10
    - Speed up development cycles: days rather than months

  ⇒ Rapid Language Adaptation from universal models
- Bridge the gap between language and technology experts
  - Technology experts do not speak all languages in question
  - Native users are not in control of the technology

Introduction & Motivation

- Many Languages – so what?
  - Growing Language Diversity on the web
  - Why do we need Speech Processing in many languages?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Language Extinction
  - Do we care? What can we do about?
- Challenges of Multilingual Speech Processing
  - Lack of Resources
  - Lack of Experts
- Solutions
  - Intelligent Learning Systems
  - Prior Work: GlobalPhone and FestVox
  - Rapid Language Adaptation Server

Prior Work: GlobalPhone and FestVox

GlobalPhone

Arabic	Croatian	Turkish
Ch-Mandarin	Czech	+ Thai
Ch-Shanghai	Portuguese	+ Creole
German	Russian	+ Polish
French	Spanish	+ Bulgarian
Japanese	Swedish	+ ... ???
Korean	Tamil	

Multilingual Database

- Widespread languages
- Native Speakers
- Uniform Data
- Broad Domain
- Large Text Resources
  - ➔ Internet, Newspaper

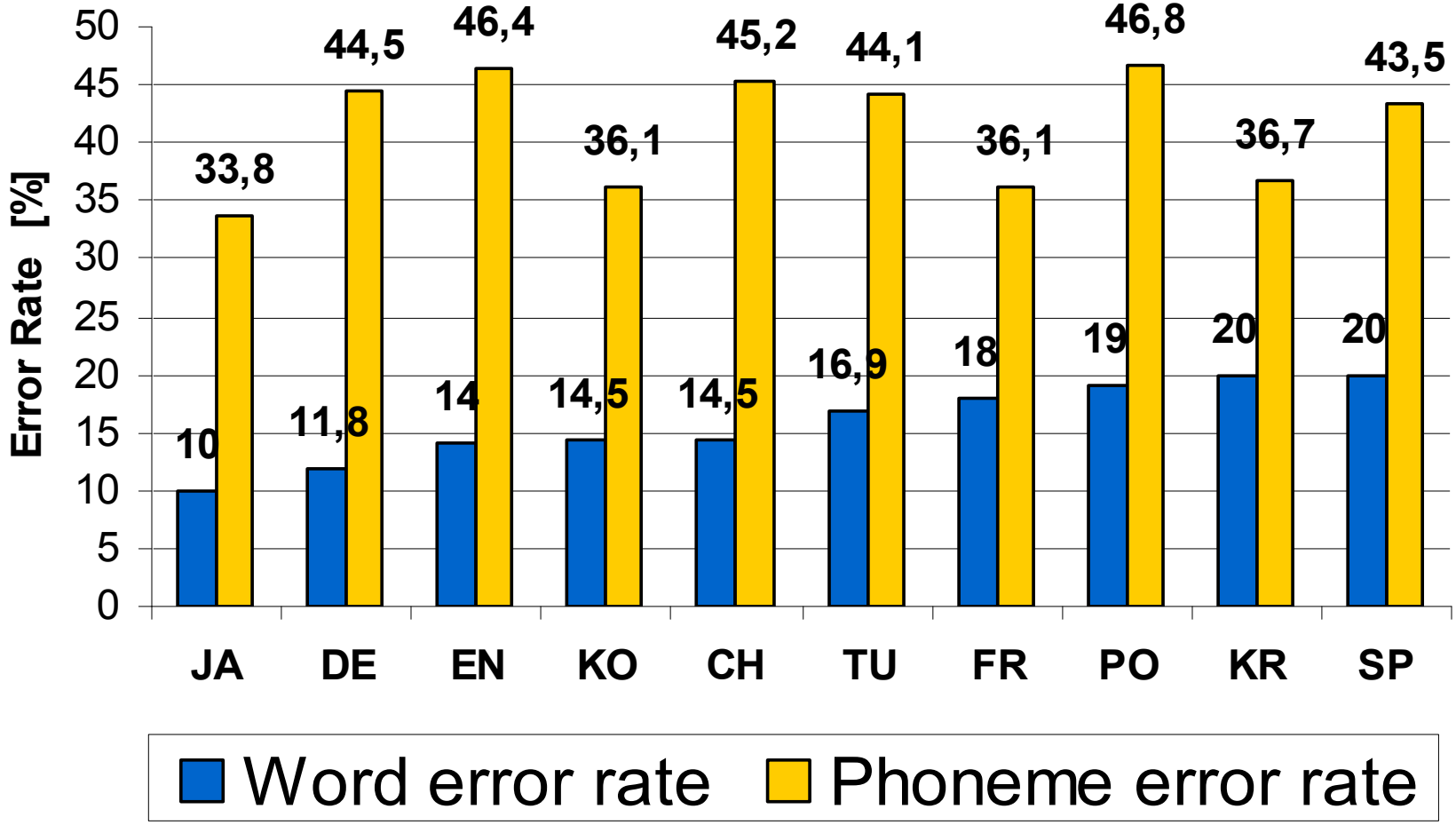
Corpus

- 19 Languages … counting
- ≥ 1800 native speakers
- ≥ 400 hrs Audio data
- Read Speech
- Filled pauses annotated

Available from ELRA

http://www.cs.cmu.edu/~tanja/GlobalPhone

GlobalPhone Recognizers in 10 Languages



Prior Work: GlobalPhone and FestVox

FestVox: Building Synthetic Voices

⇒ http://festvox.org [Black and Lenzo 2000]

- Documentation, Tools, Scripts, Examples
  - Building Synthetic Voices in the Festival Speech Synthesis System
- Supports:
  - Diphone, unit selection, (later Statistical Parametric Synthesis)
  - Lexicon, letter to sound rules
  - Text processing support.
- Example Languages:
  - CMU development: Croatian, Thai, Chinese (Mandarin), Japanese, Catalan, Spanish, Nepali
  - Non-CMU: Italian, Malay, Maori, Mongolian, Spanish, Telugu, Hindi, Japanese, English (Many), German, Swedish, Polish, …

TTS Build Tasks

- Tasks:
  - Define phone set, pronunciations (LTS vs. Lexicon)
  - Design prompt list, Record data
  - Write text front-end (Number, symbol expansion)
  - Write/train prosody model
  - Deal with peculiarities (segmentation, no vowels, etc.)

- Results strongly correlated to effort
  - Must-have for funded project
  - Involve speech experts
  - Almost random distribution rights
    - Others can't always use the previous results
    - No explicit copyrights (and no way to change them)
  - Results often not in format for re-use

Joint Speech Model Development

CMU projects: Arabic, Thai, Croatian, Farsi

- Shared audio data collection
  - Prompts with phonetic coverage
  - Lots of (ASR) / Single (TTS) speaker(s)
- Shared Phone set
  - Sometimes “similar” e.g. with/without Tone
- Shared Pronunciation Data
  - (Note) input and output are different vocab

But we need a much tighter coupling ….

Introduction & Motivation

- Many Languages – so what?
  - Growing Language Diversity on the web
  - Why do we need Speech Processing in many languages?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Language Extinction
  - Do we care? What can we do about?
- Challenges of Multilingual Speech Processing
  - Lack of Resources
  - Lack of Experts
- Solutions
  - Intelligent Learning Systems
  - Prior Work: GlobalPhone and FestVox
  - Rapid Language Adaptation Server

SPICE

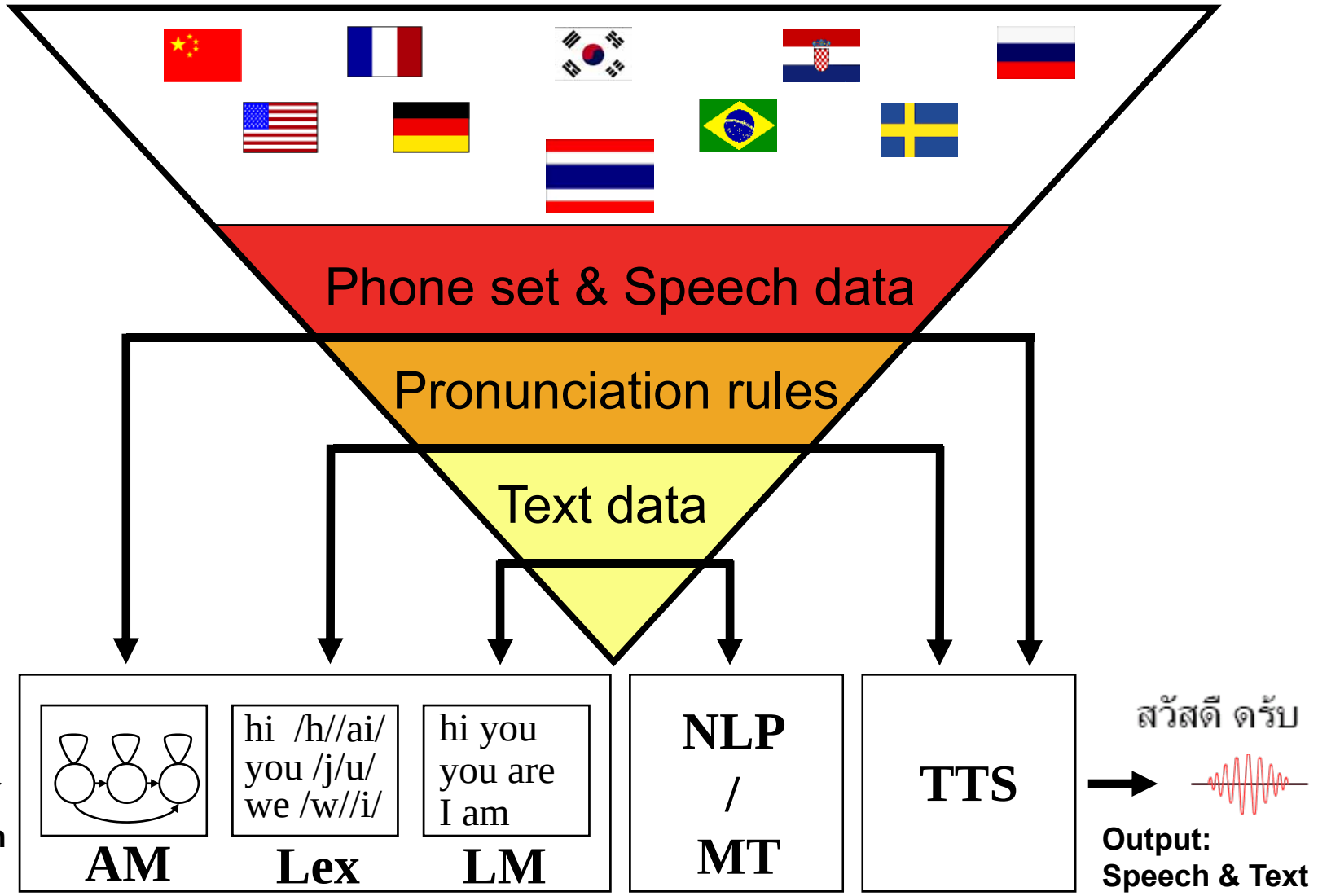
Speech **P**rocessing:

Interactive **C**reation and **E**valuation toolkit

- National Science Foundation, Grant 10/2004, 3 years
- Principle Investigator Tanja Schultz, Alan Black
- Bridge the gap between technology experts → language experts
  - Automatic Speech Recognition (ASR),
  - Machine Translation (MT),
  - Text-to-Speech (TTS)

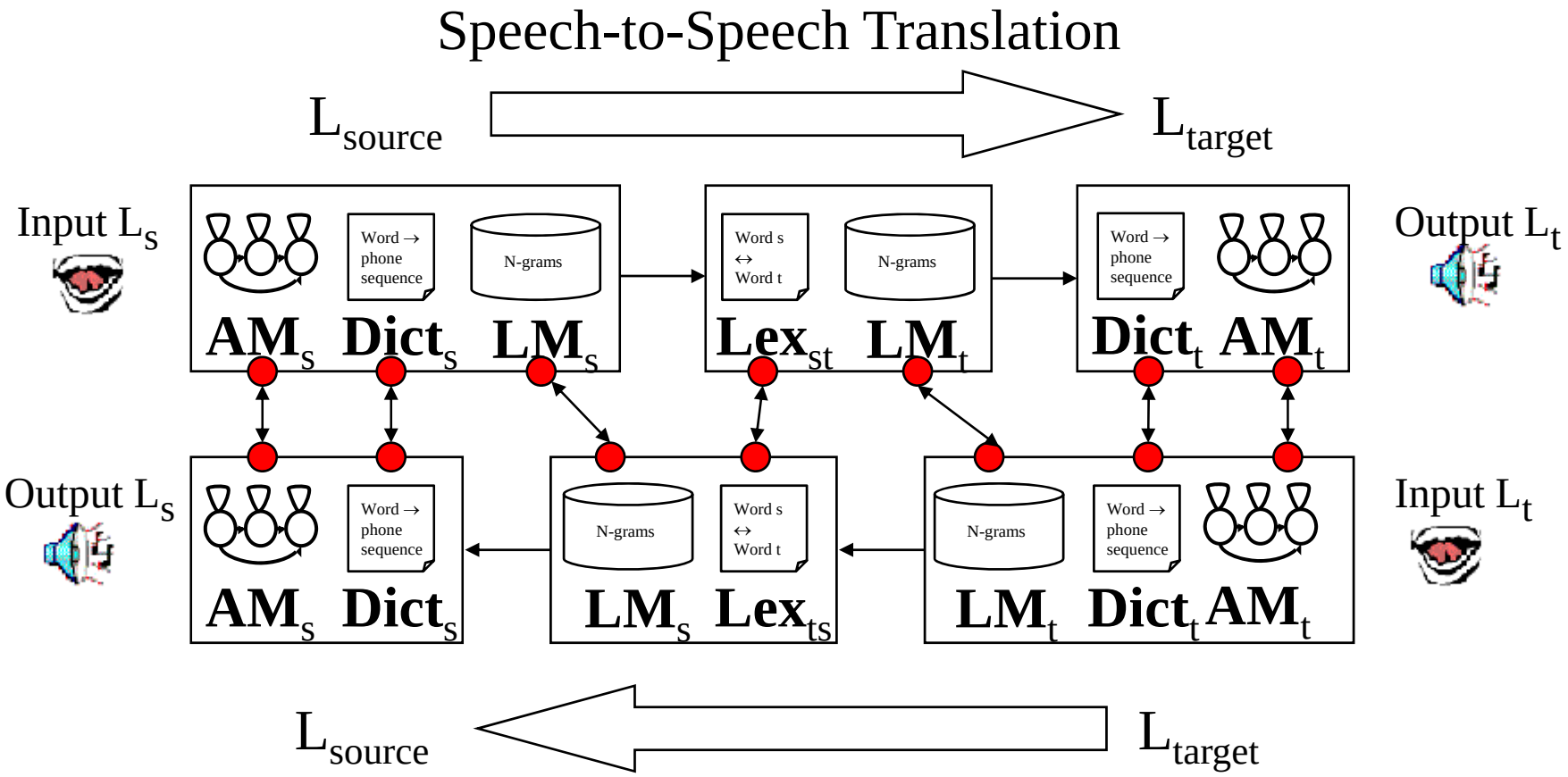
- Develop web-based intelligent systems
  - Interactive Learning with user in the loop
  - Rapid Adaptation of universal models to unseen languages
- SPICE webpage *http://cmuspice.org*

Speech Processing Systems



SPICE Design Principles

1. Data Sharing across Languages
   → Language Universal Models
2. Knowledge Sharing across System Components



SPICE: Building Support for a new Language

- SPICE gathers and archives:
  - Appropriate text data
  - Appropriate audio data
- SPICE solicits and defines:
  - Phoneme set
  - Rich prompt set
  - Lexical pronunciations
- … produces:
  - Pronunciation model
  - ASR acoustic model
  - ASR language model
  - TTS voice
- … maintains and documents:
  - Projects and users login
  - Data and Models

Tutorial Outline

- Introduction and Motivation
  - Motivation
  - History and Leveraged Work
  - Rapid Language Adaptation Server: Spice
- SPICE in detail
  - Text collection & Prompt Selection
  - Phone set specification, Lexical construction
  - ASR Bootstrap & training
  - Language model, TTS Voice building
  - Testing and Tuning
- Latest Experiments and Results
- Lessons Learnt from past studies
- Future

CMU SPICE

Welcome to SPICE

Getting started

SPICE is a web-based system for building an end-to-end speech system (including Automatic Speech Recognition and Text-To-Speech) in your own language.

Existing Users

Login with your account:

Login tanja

Password **********

Login

New Users

Create a new account:

Login

Password

Re-type Password

Email

Create new account

Login and Project Registration

- Separate “projects“ for each language
  - could share info between different projects
- All tasks times are logged
  - Allow us to do cost/efficiency studies

CMU SPICE - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)

build language model first

- Lexicon pronunciation creation (help)

build language model first

- Build acoustic model (help)
- Build language model (help)
- Create ASR system
- Create speech synthesis voice

User: **TanjaSchultz** Language: **Klingon** Project: **Interspeech2007** [Logout]

SPICE Project

You must do the following to build support for your language:

- Text collection and selection
- Audio collection
- Phoneme set specification
- Lexicon pronunciation creation
- Speech recognition acoustic model creation
- Speech recognition language model creation
- Speech synthesis voice creation

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
  select prompts first
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)

  select phonemes first
  build language model first
- Lexicon pronunciation creation (help)
  select phonemes first
  build language model first
- Build acoustic model (help)
- Build language model (help)
- Create ASR system
- Create speech synthesis voice

User: **TanjaSchultz** Language: **Klingon** Project: **Interspeech2007** [Logout]

Text collection and selection

Text Directory: /data/www/html/Spice/spice/applications/TanjaSchultz/Klingon/Interspeech2007/text
Action:

Obtain corpus
You can either upload a corpus directly, or crawl the web for one.

- Crawl the Internet: Enter URL: [ ] Depth: (1) ( ) 2 Crawl

  25% - concatenating
  (view crawl log)
  After clicking "Crawl," the crawl will run on its own in the background. When it has finished, the symbol next to it will turn green.
- [ ] Browse... Upload_Text

  Note that the text file uploaded must be **uncompressed**, and '**plain text**' (i.e. not a PDF, Word document, etc.)

Obtain prompts
You can upload your own prompts, or use find_prompts to have SPICE automatically generate prompts.

- [ ] Browse... Upload_Prompts

  Note that the text file uploaded must be **uncompressed**, and '**plain text**' (i.e. not a PDF, Word document, etc.)
- find_prompts Find nice prompts ☐ Check to allow all characters in prompts, leaving text 'as is'

  After clicking "find_prompts," the prompt finder will run on its own in the background. When it has finished, the symbol next to it will turn green.

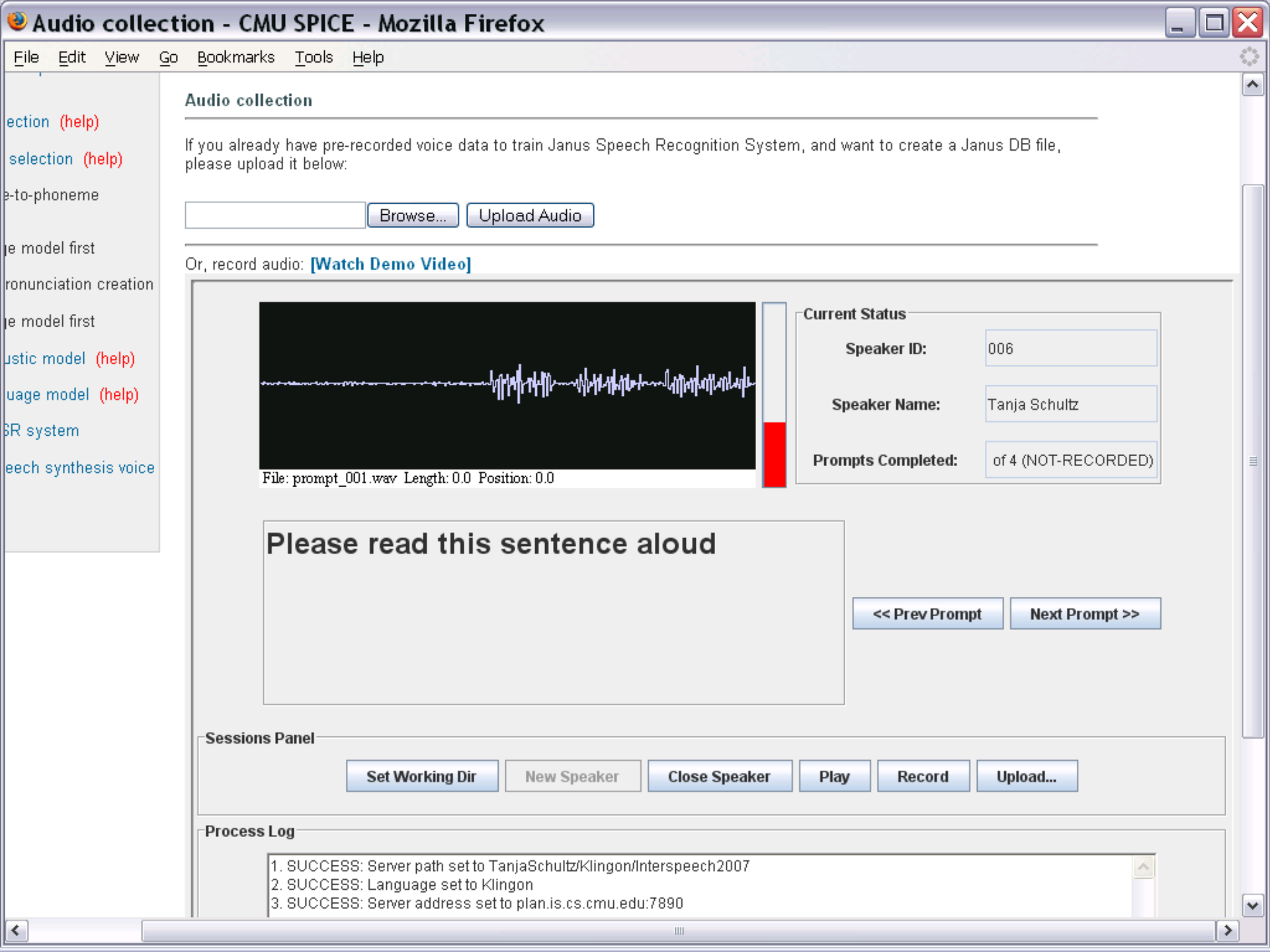
Text Collection

- We need text data for the target language
  - Web crawler
  - Plus boost data from similar sites
- Language encoding
  - Non-trivial, but ...
  - Deal with very common alphabets
  - Internally all utf-8
- In-domain vs. general text
- Character analysis
  - Find the character classes:
  - casing, numerals, punctuation etc

Prompt Selection

- Prompts for recording:
  - Collection without transcription
  - “Good” coverage higher chance to give “clean” models

- Prompts should be:
  - Easy to say (no hard words, numerals etc)
    - Contain high frequency words
    - Easy to say in one breath group, 5-15 words
  - “Phonetically” rich / rich in variability
    - But we have no phonetic information yet
    - Make them orthographically rich
    - Greedily select to maximize tri-graphs


Audio collection - CMU SPICE - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
ection (help)
selection (help)
e-to-phoneme
e model first
ronunciation creation
e model first
stic model (help)
uage model (help)
R system
eech synthesis voice
Audio collection
If you already have pre-recorded voice data to train Janus Speech Recognition System, and want to create a Janus DB file, please upload it below:
Browse...
Upload Audio
Or, record audio: [Watch Demo Video]
File: prompt_001.wav Length: 0.0 Position: 0.0
Current Status
Speaker ID: 006
Speaker Name: Tanja Schultz
Prompts Completed: of 4 (NOT-RECORDED)
Please read this sentence aloud
<< Prev Prompt
Next Prompt >>
Sessions Panel
Set Working Dir
New Speaker
Close Speaker
Play
Record
Upload...
Process Log
1. SUCCESS: Server path set to TanjaSchultz/Klingon/Interspeech2007
2. SUCCESS: Language set to Klingon
3. SUCCESS: Server address set to plan.is.cs.cmu.edu:7890

Speech Data Collection

- Online Audio Recording Tool
  - Collaboratively record large number of speakers
  - Speakers may be separate from developer
- Visual feedback during recording
- Automatic upload on completion
- Java based for portability
  - Works with *many* browsers
- In control of recording
  - We can control the recording format
  - File contents and directory structure

Phoneme Selection - CMU SPICE - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help) select prompts first
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)

select phonemes first
build language model first

- Lexicon pronunciation creation (help)

select phonemes first
build language model first

- Build acoustic model (help)
- Build language model (help)
- Create ASR system
- Create speech synthesis voice

User: **TanjaSchultz** Language: **Klingon** Project: **Interspeech2007** [Logout]

Phoneme set specification

This is a tool which will display all IPA phoneme. As a naive user, you can choose and give names to phonemes you wish your Speech Engine to use. After you have finished, you can click the "Submit" button to create the new acoustic model on the fly.

Consonants (Pulmonic): Please choose the consonant sounds you'd like to have in your new acoustic models by giving it a name in the textbox next to it.

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	[P] p [ ] pʲ [B] b [ ] bʲ		[T] t [ ] tʲ [D] d [ ] dʲ			[ ] ʈ [ ] ɖ	[C] c [ ] ɟ	[K] k [ ] kʲ [G] g [ ] gʲ	[ ] q [ ] ɢ		[ ] ʔ
Nasal	[M] m [ ] mʲ	[ ] ɱ	[N] n [Nj] nʲ			[ ] ɳ	[ ] ɲ	[ ] ŋ	[ ] ɴ		
Trill	[ ] ʙ		[ ] r						[ ] ʀ		

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
  select prompts first
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)
  build language model first
- Lexicon pronunciation creation (help)
  build language model first
- Build acoustic model (help)
- Build language model (help)
- Create ASR system
- Create speech synthesis voice

User: **TanjaSchultz** Language: **Klingon** Project: **Interspeech2007** [Logout]

Phoneme selection

Result: Phoneme set saved in settings
Mapfile saved in KlingonmapFile

Phonemes selected:
P
B
T
D
C
K
G
M
N
I
U
E
O
A
Nj

Phoneme Selection

- Selection from standard IPA chart
- User's names for phonemes
  - Can match user's lexicon (if one exists)
  - Can match user's familiarity
- Audio feedback
  - Click to hear recording of each phone
- Allows us to map user's phone names
  - We map phones to IPA
  - Get phonetic features for user's phones
  - (what are vowels, what are stops etc)
- Bootstrap from Multilingual Phone Sets
  (see Acoustic Model Initialization)

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)
- Lexicon pronunciation creation (help)
- Build acoustic model (help)
- Build language model (help)
- Create ASR system

Phoneme labels for your language:
P B T D C K G M N I U E O A NJ

User: **TanjaSchultz** Language: **Klingon** Project: **Interspeech2007** [Logout]

Lexicon pronunciation creation

Initial Grapheme to Phoneme Rules

Please input an initial Grapheme to Phoneme (G2P) rule of your language.

Based on this rule, our system will "guess" the correct pronunciationof words in your language. You are able to view the predicted pronunciation,change it, delete it, or type a correct pronunciation for this word.The correct pronunciation will be saved into your dictionary and oursystem will make use of this information to make a better "guess"in predicting pronunciation of new words.

Now please type in Grapheme to Phoneme rule (G2P) for us. Just typeone of the most common pronunciation for each grapheme. Thanks.

Upload g2p [Browse...] Upload char.info [Browse...] [Upload]

[ ] ○ uppercase ○ lowercase ○ punctuation mark ○ number ○ others

[ ] ○ uppercase ○ lowercase ○ punctuation mark ○ number ○ others

. [ ] ○ uppercase ○ lowercase ◉ punctuation mark ○ number ○ others

C [C] ◉ uppercase ○ lowercase ○ punctuation mark ○ number ○ others

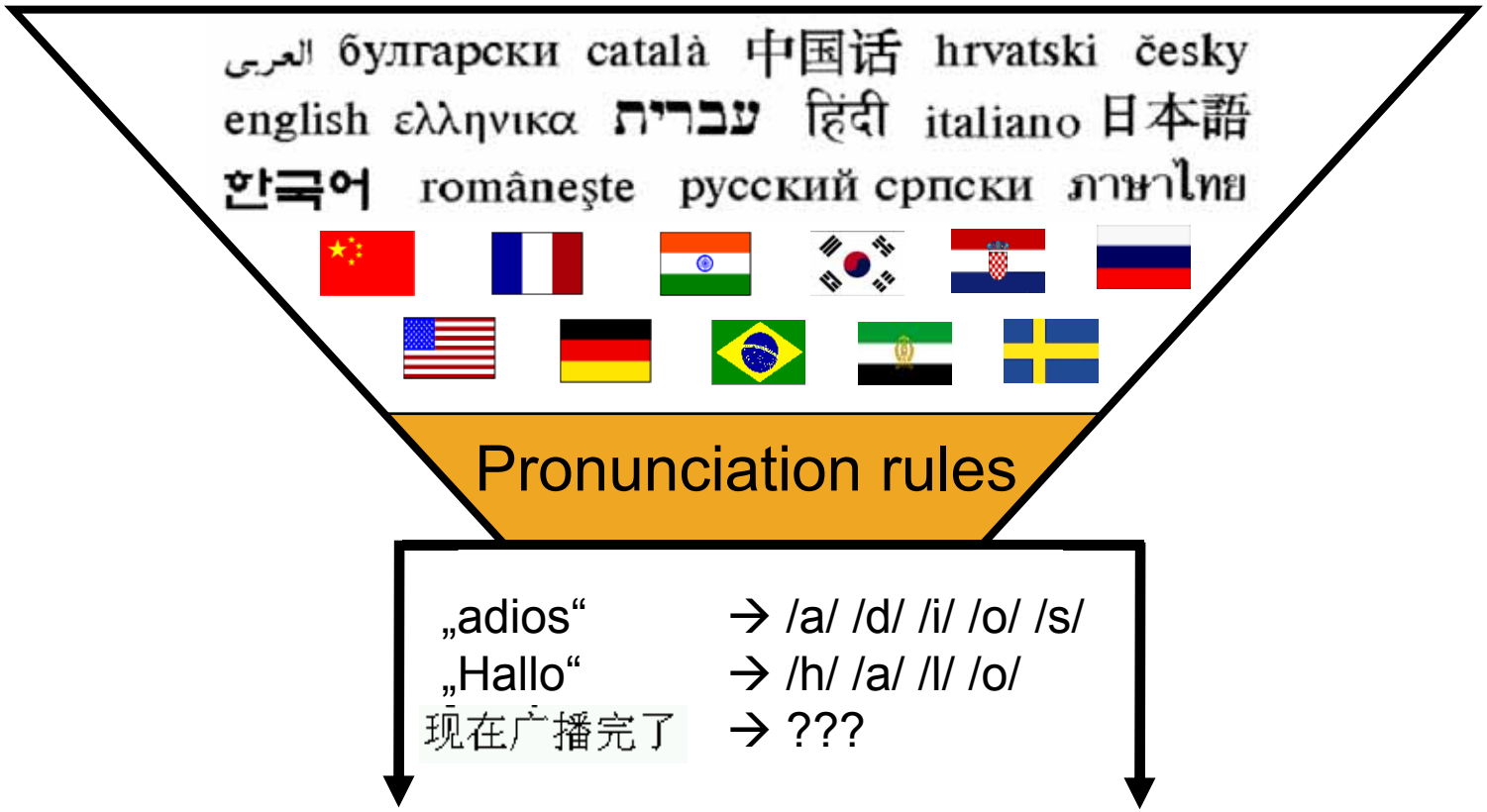
E [E] ◉ uppercase ○ lowercase ○ punctuation mark ○ number ○ others

I [I] ◉ uppercase ○ lowercase ○ punctuation mark ○ number ○ others

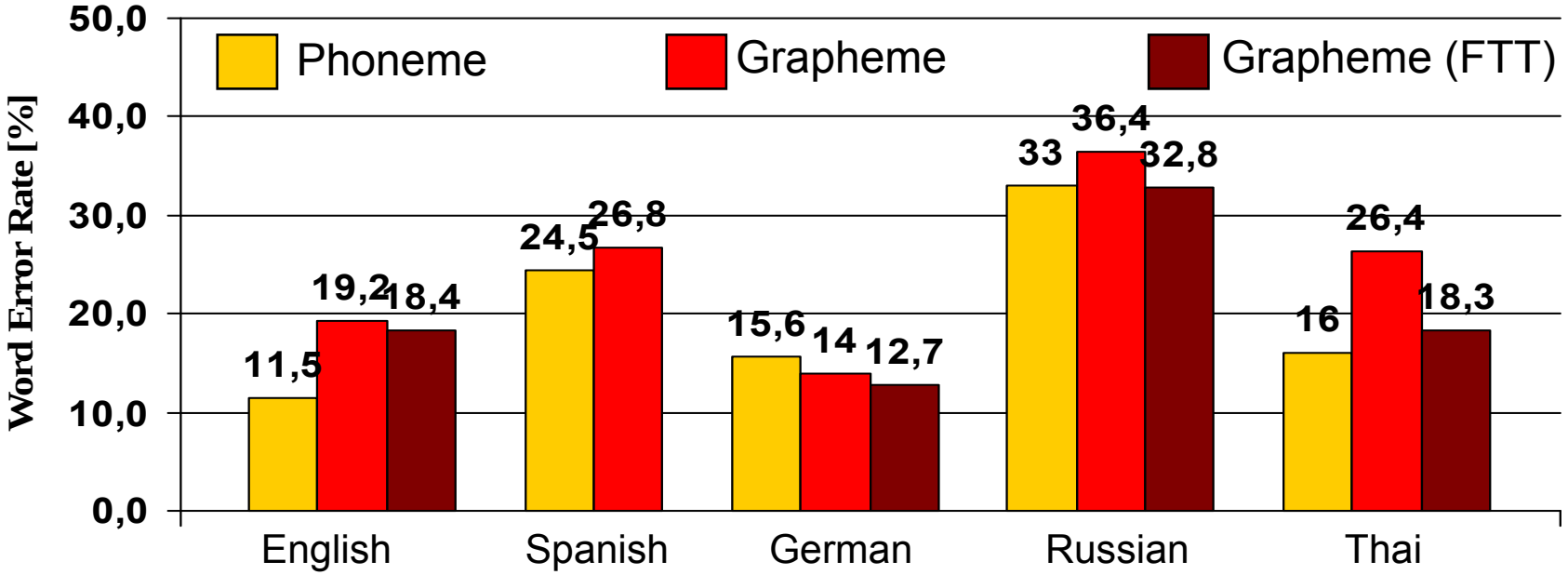
M [M] ◉ uppercase ○ lowercase ○ punctuation mark ○ number ○ others

P [ ] ○ uppercase ○ lowercase ○ punctuation mark ○ number ○ others

Pronunciation Dictionary / Lexicon



Phoneme- vs Grapheme-based ASR



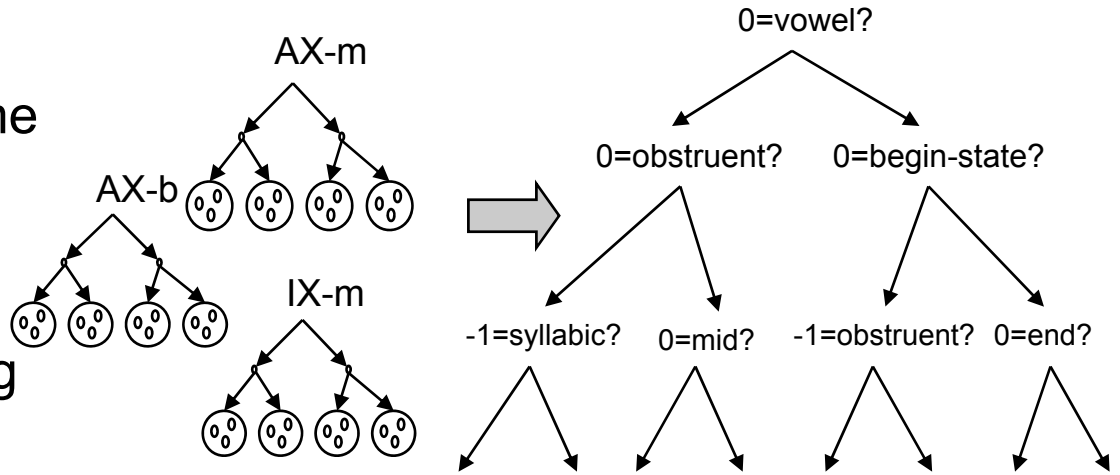
Problem:

- 1 Grapheme ≠ 1 Phoneme

Flexible Tree Tying (FTT):

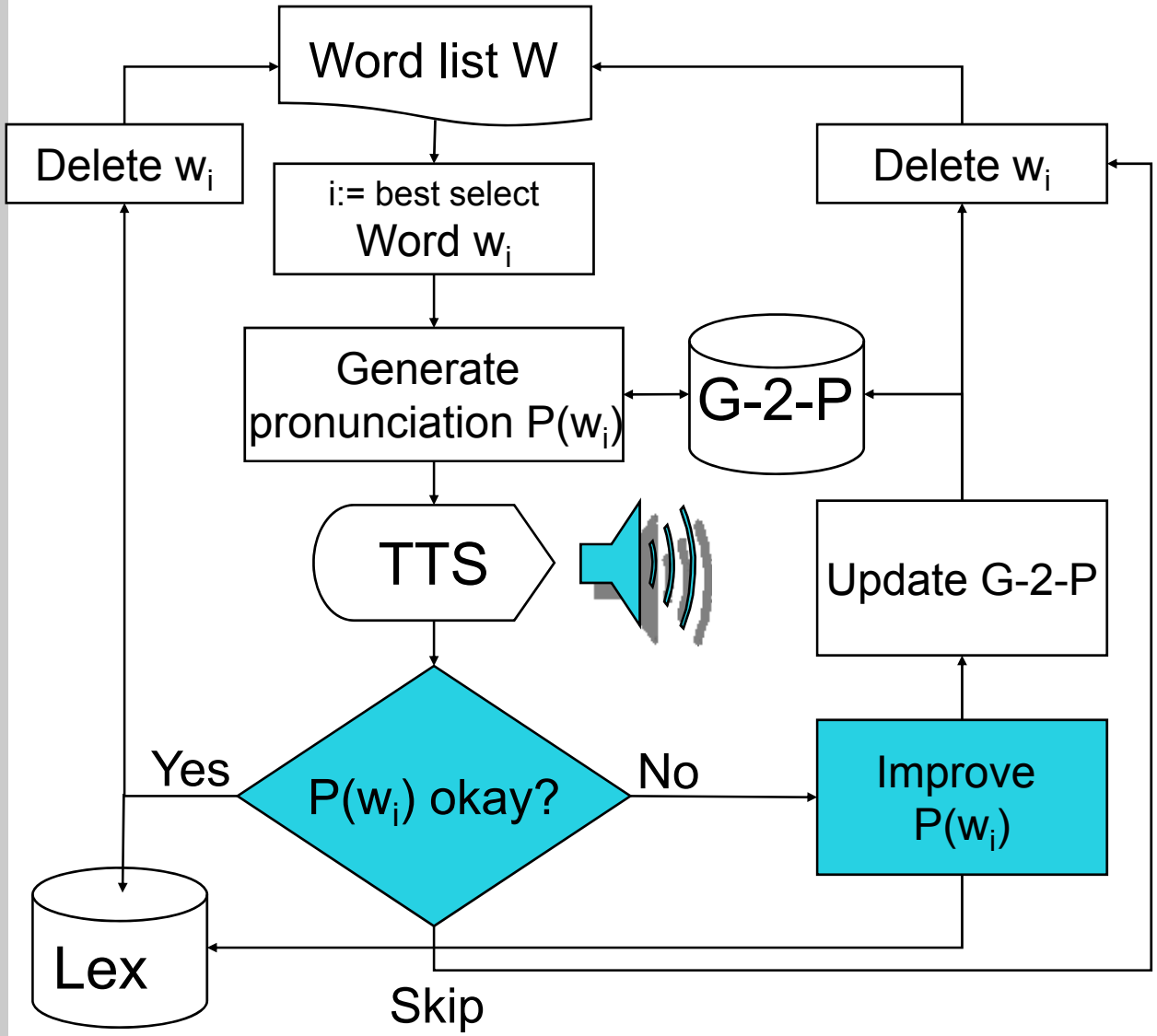
One decision tree

- Improved parameter tying
- Less over specification
- Fewer inconsistencies



Carnegie Mellon

Dictionary: Interactive Learning



* Follow the work of Davel&Barnard

* Word list: extract from text

* G-2-P
- explicit map rules
- neural networks
- decision trees
- instance learning (grapheme context)

* Update after each w_i
→ effective training

User

Lexicon - CMU SPICE - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)
- Lexicon pronunciation creation (help)
- Build acoustic model (help)
- Build language model (help)
- Create ASR system

voice

Phoneme labels for your language:

P B T D C K G M N I U E O A NJ

User: **TanjaSchultz** Language: **Klingon** Project: **Interspeech2007** [Logout]

Lexicon pronunciation creation

Rule entry

28.571428571429% Finished

new word:

to

system suggested pronunciation: T O listen to it Accept Pronunciation

If you want to skip this word and work on it later, please click Skip this word

If you don't think it's a valid word in your language, please click Remove this word

To save your work to resume later and build a lexicon with your current input, click Pause and Build Lexicon

Issues and Challenges

- How to make best use of the human?
  - Definition of successful completion
  - Which words to present in what order
  - How to be robust against mistakes
  - Feedback that keeps users motivated to continue

- How many words to be solicited?
  - G2P complexity depends on the language (SP easy, EN hard)
  - 80% coverage
    hundred (SP) to thousands (EN)
  - G2P rule system perplexity

Language	Perplexity
English	50.11
Dutch	16.80
German	16.70
Afrikaans	11.48
Italian	3.52
Spanish	1.21

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
  select prompts first
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)

  build language model first
- Lexicon pronunciation creation (help)
  build language model first
- Build acoustic model (help)
- Build language model (help)
- Create ASR system
- Create speech synthesis voice

User: **TanjaSchultz** Language: **Klingon** Project: **Interspeech2007** [Logout]

Build acoustic model

1. Janus Database Creation

- Create Janus Database

2. Acoustic Model Initialization

- Initialization: IN PROGRESS...
- Initialize_Acoustic_Models

3. Train Context Independent Acoustic Models

- Computing Labels: IN PROGRESS...
- Computing Cepstral Means: NOT STARTED YET.
- Computing LDA matrix: NOT STARTED YET.
- K-Means Clustering over the codebooks: NOT STARTED YET.
- EM Training: NOT STARTED YET.
- Train_CI_Models

4. Prepare for Training Context Dependent Acoustic Models

- Making Polyphone Trees: IN PROGRESS...
- EM Training over the code-books: NOT STARTED YET.
- Clustering Contexts: NOT STARTED YET.
- Performing the Splits: NOT STARTED YET.
- Prep_CD_Models

5. Train Context Dependent Acoustic Models

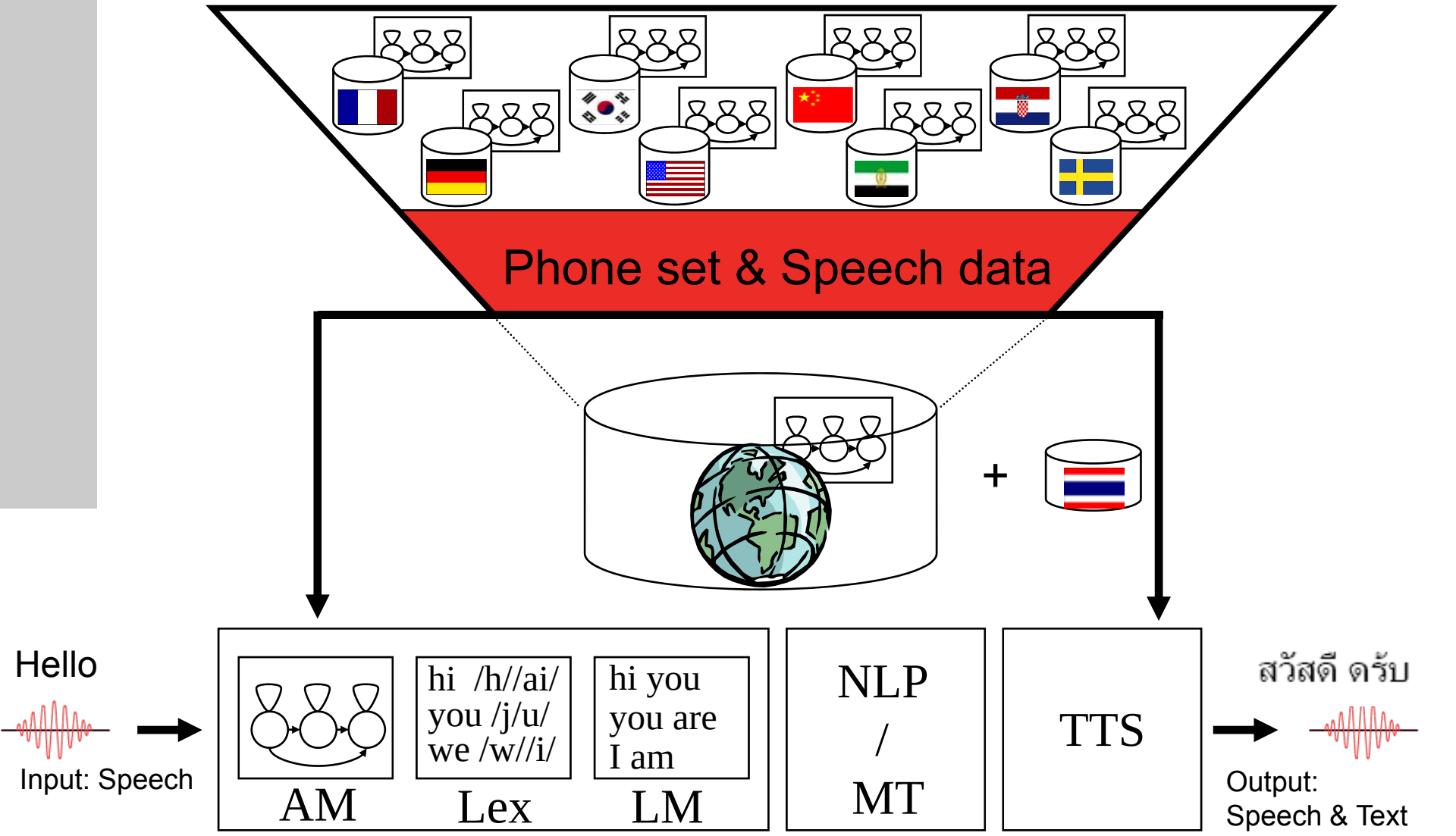
- Context Dependent Acoustic Models Training: IN PROGRESS...
- Train_CD_Models

Acoustic Model Building

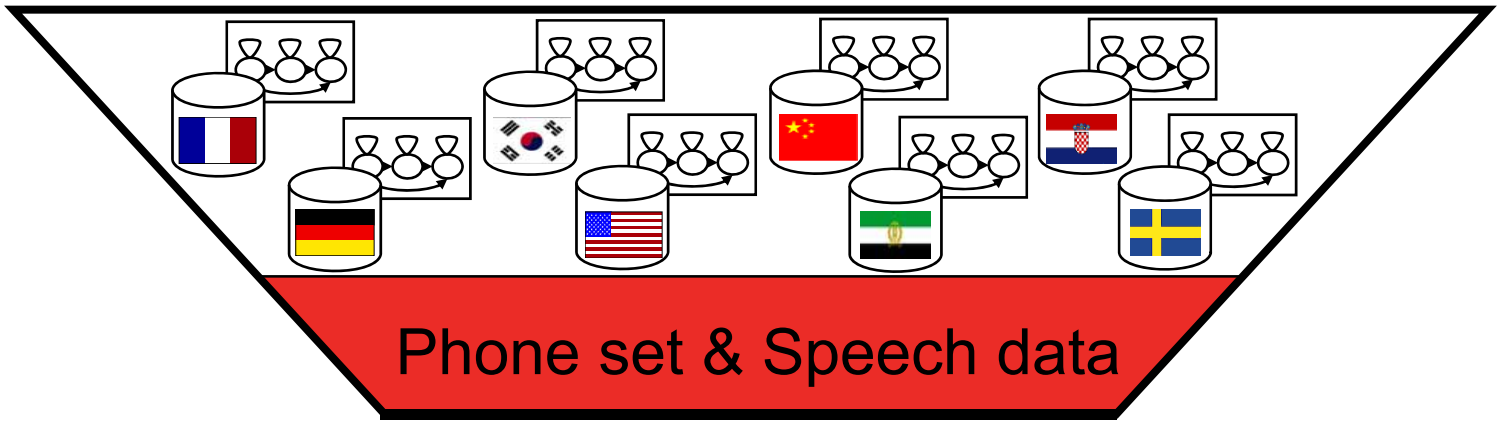
- Acoustic Model Building requires:
  - Transcribed Speech Data
  - Phone set definition
  - Pronunciation Lexicon
    (if transcripts are on word level only – standard case)

- Two step process:
  1. Model Initialization
  2. Model Training

Acoustic Model Initialization



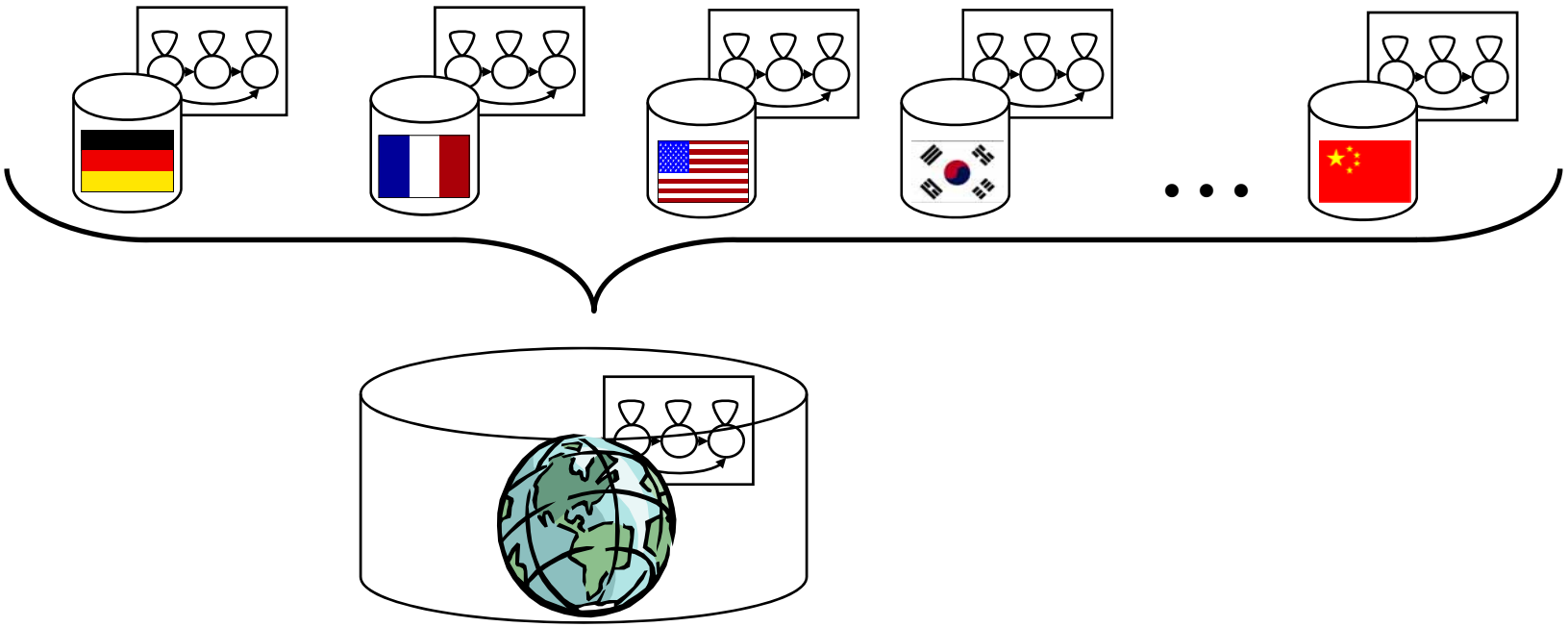
Acoustic Models: Multilingual Data



Step 1:

- Uniform multilingual database (GlobalPhone)
- Build Monolingual acoustic models in many languages

Multilingual Acoustic Modeling



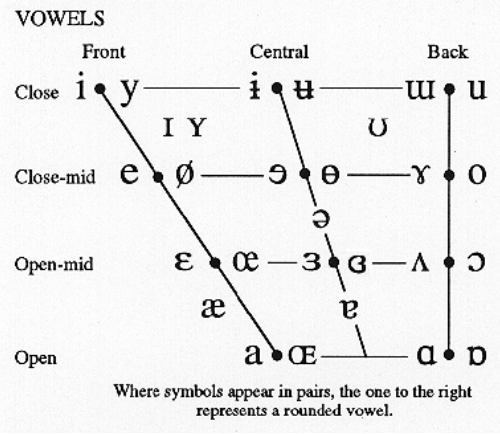
Step 2:

- Combine monolingual acoustic models to a set of multilingual “language independent” acoustic model

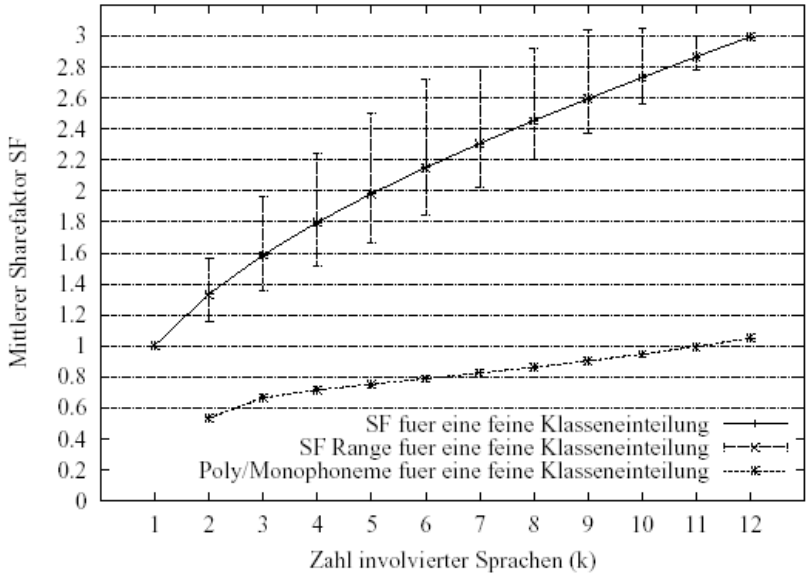
Universal Sound Inventory

Speech Production is independent from Language ⇒ IPA

1) IPA-based Universal Sound Inventory



2) Each sound class is trained by data sharing

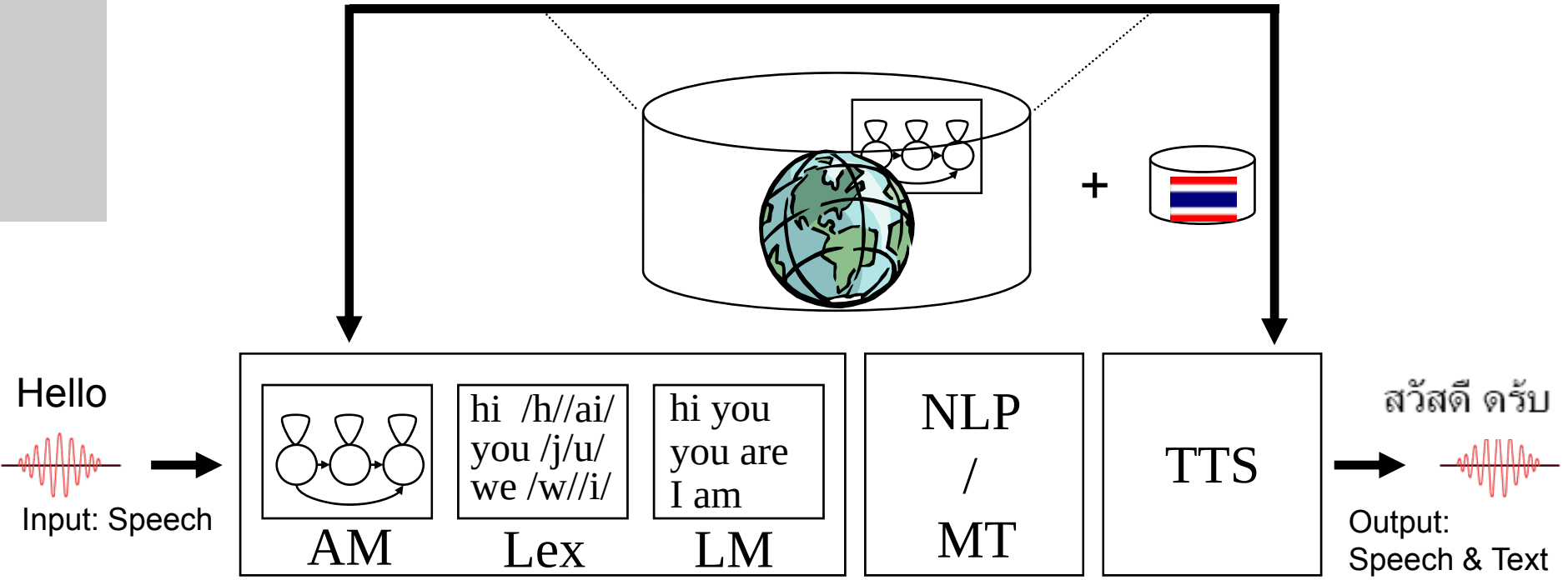


- Reduction from 485 to 162 sound classes
- *m,n,s,l* appear in all 12 languages
- *p,b,t,d,k,g,f* and *i,u,e,a,o* in almost all

Cross-language Bootstrapping

Step 3:

- Define mapping between ML set and new language
- Bootstrap acoustic model of unseen language



Acoustic Model Building - Configuration

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)
- Lexicon pronunciation creation (help)
- Build acoustic model (help)
- Build language model (help)
- Test ASR system
- Create speech synthesis voice

User: **demo** Language: **eng** Project: **walk_dec3** [Logout]

Build acoustic model

Step 1. Janus Database Creation

- Configure_Training

ASR build configured successfully View/Hide Results >>
According to our analysis, you have recorded speech data from 3 speakers amounting to a total of 12.0 minutes.

Total number of iterations to be performed during Step 2 are 5.

1. Iterations 1 - 3 : Leave-one-out testing where 2 speakers data is used for training and 1 speaker data is used for testing the acoustic models during each round.
2. Iteration 4: 90% of each speaker's utterances are used for training and the rest of the utterances are used for testing the acoustic models.
3. Iteration 5: All the recorded data of the 3 speakers is used for training the acoustic models.

Acoustic Model Building - Configuration

- Checks dependencies and errors
  - Lexicon and phone set correspond
  - Words in recorded prompts are covered by the lexicon
- Divides the recorded data into training and test sets
- Performance evaluation
  - Few data: K-fold cross-validation, with K = #speakers
  - More data: Data split into 90% (train) and 10% (test)

Acoustic Model Building - Training

- Audio collection (help)
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)
- Lexicon pronunciation creation (help)
- Build acoustic model (help)
- Build language model (help)
- Test ASR system
- Create speech synthesis voice

Step 2. Acoustic Model Training

- Train_Acoustic_Models LOG_FILE

- Initialization: COMPLETED!
- Computing Labels: COMPLETED!
- Computing Cepstral Means: COMPLETED!
- Computing LDA matrix: COMPLETED!
- K-Means Clustering over the codebooks: COMPLETED!
- EM Training: COMPLETED!
- Segmentation of Speech: COMPLETED!
- Making Polyphone Trees: COMPLETED!
- EM Training over the code-books: COMPLETED!
- Clustering Contexts: COMPLETED!
- Performing the Splits: COMPLETED!
- Context Dependent Acoustic Models Training: COMPLETED!

View/Hide Results >>

Iteration: 1
Utterances = 191 Substitutions = 52 Insertions = 64 Deletions = 70 Total Words = 1515 wer = 12.2772277228%

Iteration: 2
Utterances = 192 Substitutions = 63 Insertions = 160 Deletions = 12 Total Words = 1523 wer = 15.4300722259%

Iteration: 3
Utterances = 187 Substitutions = 36 Insertions = 14 Deletions = 28 Total Words = 1471 wer = 5.30251529572%

Iteration: 4
Utterances = 59 Substitutions = 28 Insertions = 19 Deletions = 15 Total Words = 565 wer = 10.9734513274%

Acoustic Model Building - Training

- Requires successful configuration
- Generalized Training Procedure
  - EM Training for Context Independent Models
    - 3-state HMM
    - Number of Gaussians per Model depends on data
  - EM Training for Context Dependent Models
    - Number of models depends on data
  - MFCC front-end, LDA
- Feedback to User:
  - Progress of training procedure
  - Detailed Log files of all processes
  - Results of performance evaluation

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)
- Lexicon pronunciation creation (help)
- Build acoustic model (help)
- Build language model (help)
- Create ASR system
- Create speech synthesis voice

User: **TanjaSchultz** Language: **Klingon** Project: **Interspeech2007** [Logout]

Language model

Build language model (View language model directory) (View language model build log file)

Language model build complete

Your language model is built. If you want to rebuild it, your old language model will be backed-up at:
applications/TanjaSchultz/Klingon/Interspeech2007/lm-backupAt20070813-135133

To build a language model, click this button:

Build Language Model

Language Model Building

Goal:

- Get as much relevant text data as possible
- Use the text data for
  - Generating recording prompts
  - Generating vocabulary lists
  - Build Language Models for ASR

Approach

1. User supplies an URL to SPICE for crawling
2. Crawler retrieves N documents (web-pages)
3. Compute the statistics (TF-IDF) from the N documents
4. Terms with highest TF-IDF score form query terms
5. Query search engine (Google) to get the URLs for the query terms
6. Crawl the URLs for the data

Synthesis - CMU SPICE - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)

build language model first

- Lexicon pronunciation creation (help)

build language model first

- Build acoustic model (help)
- Build language model (help)
- Create ASR system
- Create speech synthesis voice

User: **TanjaSchultz** Language: **Klingon** Project: **Interspeech2007** [Logout]

Building synthesis voice

Tasks

Voice Name: `cmu_spice_Klingon_Interspeech2007`
Voice Directory: cmu_spice_Klingon_Interspeech2007
Tasks:

- recreate voice (and delete current one)

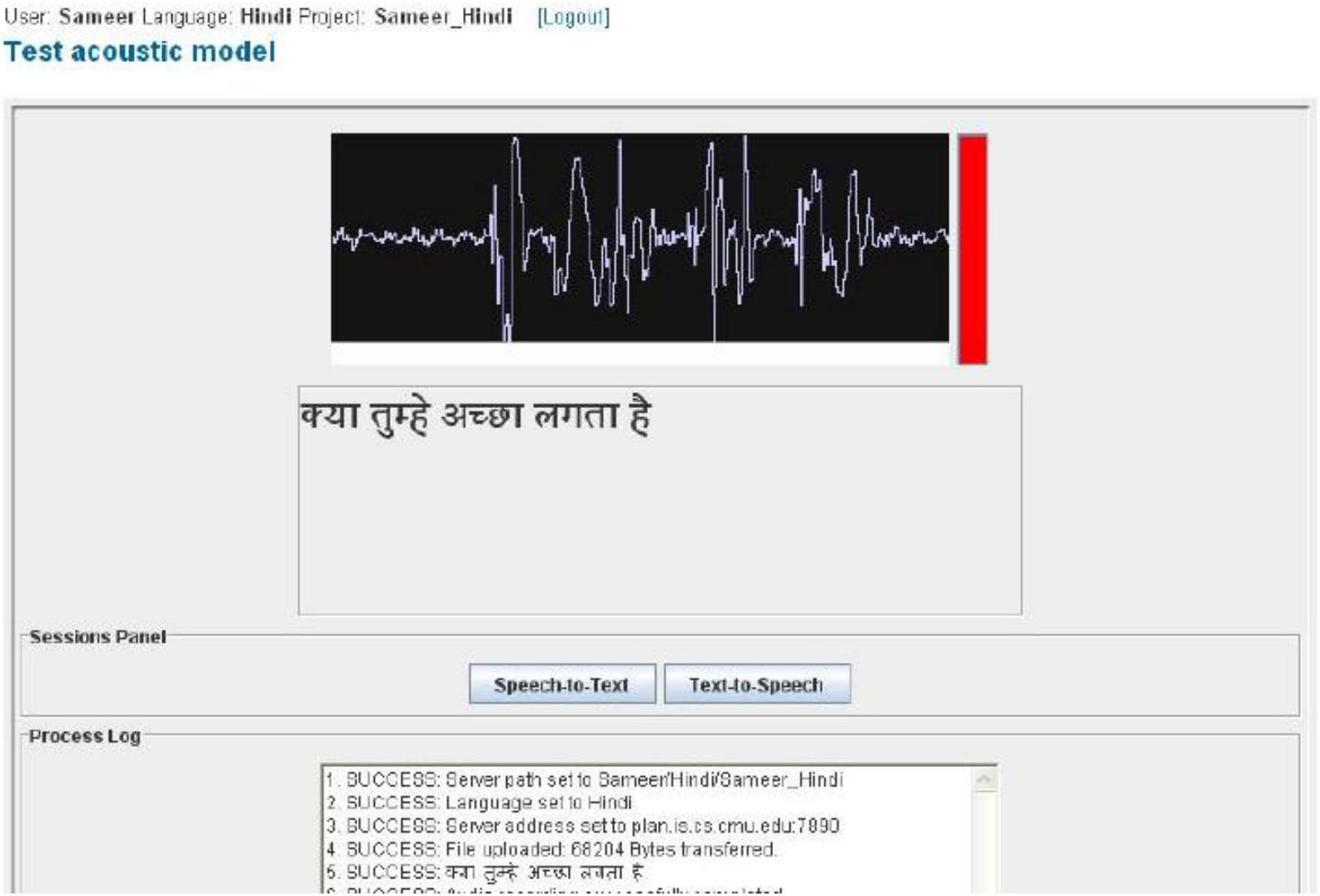
`cmu_spice_Klingon_Interspeech2007`

- import_waves `waves`
- import_prompts `txt.done.data`
- import_lexicon `lexicon lexrules`
- label_segments
- extract_params no labels
- build_models build Spectral and F0 models
- build_dur build duration model
- test_voice test the voice
- package_voice package the voice

Statistical Parametric TTS

- Text-to-speech for Applications, Common technologies:
  - Diphone: too hard to record and label
  - Unit selection: too much to record and label
- Statistical Parametric Synthesis: “just right”
  - “HMM synthesis”: ***clustergen*** trajectory synthesis
    - Clusters representing context-dependent allophones
    - Works robustly with as little as 10min speech data
    - But … Signal may sound “buzzy”, can lack varied prosody
- Voice Building Process
  - Collect 300-500 utterances from single speaker, rich prompt set
  - Lexical coverage (from Lex Learner)
  - Automatic labeling from acoustic models
  - Automatic: spectral and prosodic models
- More details tomorrow:
  John Kominek: “Synthesizer Voice Quality of New Languages”

Testing ASR-TTS



Simple echo back testing function

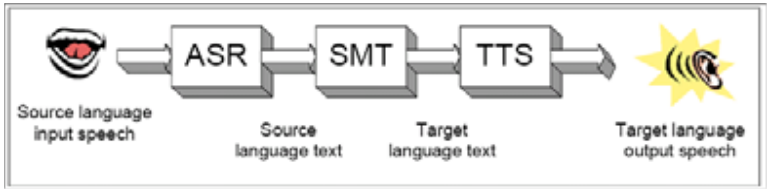
SPICE: Demo Tape

Tutorial Outline

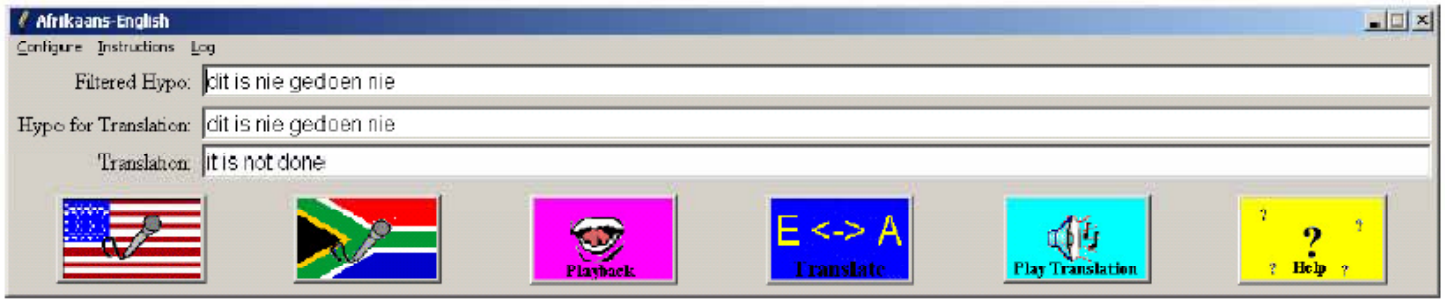
- Introduction and Motivation
  - Motivation
  - History and Leveraged Work
  - Rapid Language Adaptation Server: Spice
- SPICE in detail
  - Text collection & Prompt Selection
  - Phone set specification, Lexical construction
  - ASR Bootstrap & training
  - Language model, TTS Voice building
  - Testing and Tuning
- Latest Experiments and Results
  - Lessons Learnt from past studies
- Future

SPICE 2005: Afrikaans – English

- Goal: Build Afrikaans – English Speech Translation System with SPICE
  - Cooperation with University Stellenbosch and ARMSCOR
  - Bilingual PhD visited CMU for 3 month
  - Afrikaans: Related to Dutch and English,
    g-2-p very close, regular grammar, simple morphology
- SPICE, all components apply statistical modeling paradigm
  - ASR: HMMs, N-gram LM (JRTk-ISL)
  - MT: Statistical MT (SMT-ISL)
  - TTS: Unit-Selection (Festival)
  - Dictionary: G-2-P rules using CART decision trees

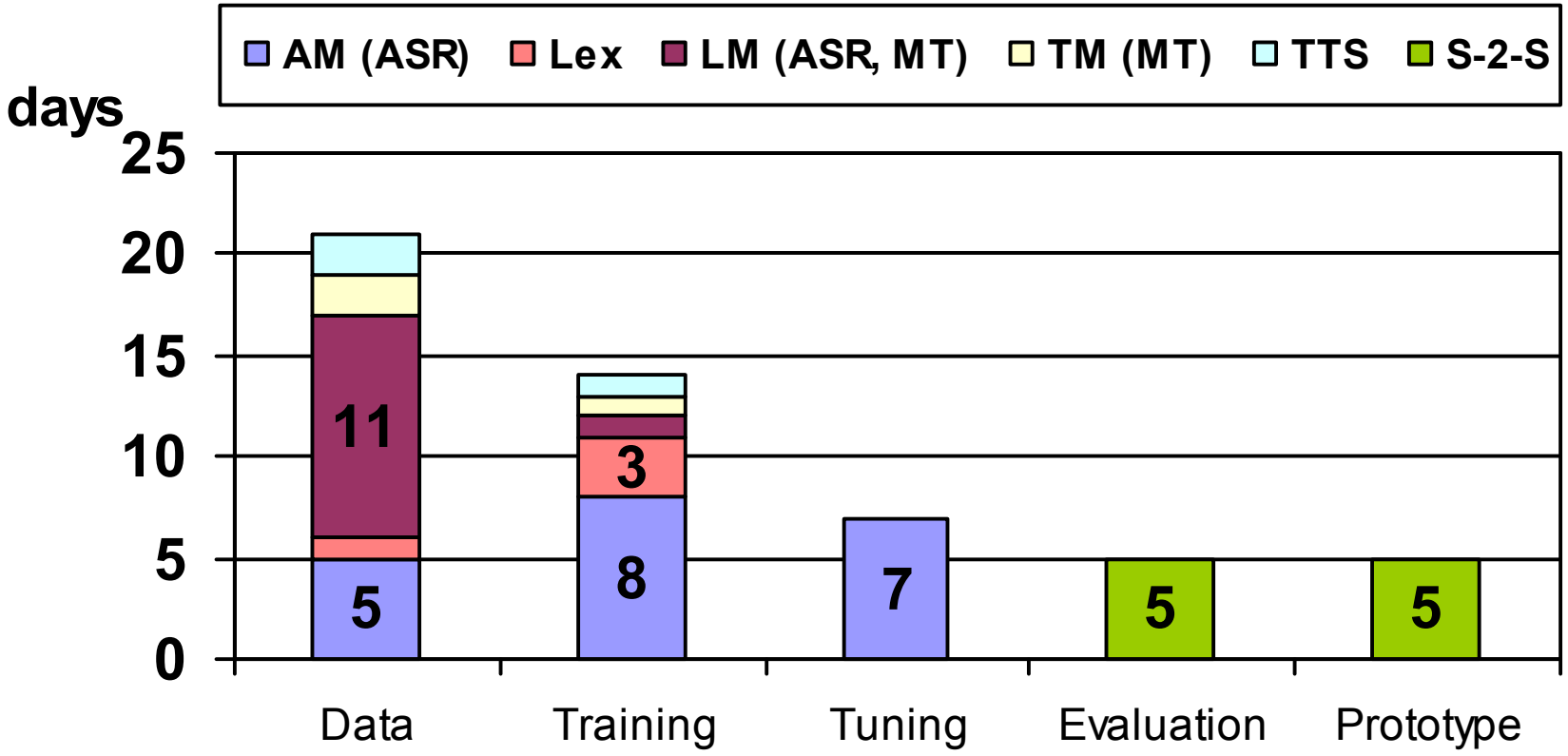


- Text: 39 hansards; 680k words; 43k bilingual aligned sentence pairs;
  Audio: 6 hours read speech; 10k utterances, telephone speech (AST)



Time Effort

- Good results: ASR 20% WER; MT A-E (E-A) Bleu 34.1 (34.7), Nist 7.6 (7.9)
- Shared pronunciation dictionaries (for ASR+TTS) and LM (for ASR+MT)
- Most time consuming process: data preparation → reduce amount of data!
- Still too much expert knowledge required (e.g. ASR parameter tuning!)



Herman Engelbrecht, Tanja Schultz, Rapid Development of an Afrikaans-English Speech-to-Speech Translator , IWSLT 2005, Pittsburgh, PA, October 2005

SPICE 2007: Field Experiments

- Now targeting *more* languages in a *shorter* time frame
- 6-weeks Hands-on Course at CMU in Spring 2007
  - Adopt native languages of participating students as targets
  - Added up to 10 different languages: Bulgarian, English, French, German, Hindi, Konkani, Mandarin, Telugu, Turkish, Vietnamese
- Teams of two students with different native language
- Course goal was to build a simple S-2-S system and use this to communicate with each other in their mother tongue
  - Solely rely on SPICE tools
  - Build speech recognition components in two languages
  - Build simple SMT component in two directions
  - Build speech synthesis components in two languages
  - Report back on problems and system shortfalls

SPICE 2007: Field Experiments (2)

- The 10 languages cover broad range of peculiarities
- Writing system:
  - Logographic Hanzi (Mandarin);
  - Cyrillic (Bulgarian);
  - Roman (German, French and English);
  - phonographic segmental (Telugu and Hindi);
  - phonographic featural (Vietnamese)
  - No script: Konkani
- Segmentation: No segmentation (Chinese); Segmentation white spaces do not necessarily indicate word (Vietnamese)
- Morphology: simple, low inflecting (English), compounding (German), agglutinating (Turkish) …
- Sound System: tonal (Mandarin and Vietnamese), stress (Bulgarian)
- G-2-P: straightforward (Turkish), challenging (Hindi), difficult (English), no relationship (Chinese), invented (Konkani)

SPICE 2007: Lessons Learnt

- It is possible to create speech processing components for 10 languages in 6-weeks using SPICE
- Each language brings new challenges
- Many SPICE features turned out to be very helpful, e.g. only ONE speaker of Konkani in Pittsburgh, web recorder allowed remote collection of more speakers

- Log: time spent in SPICE interface
- Improve interface using breakdown
- Use feedback
- Interface allows for collaborative work

Task	Time Spent [hh:mm]
Text Collection	8:35
Audio Collection	10:07
Phoneme Selection	4:05
LM building	1:25
G-2-P specs	1:30

Tanja Schultz, Alan W Black, Sameer Badaskar, Matthew Hornyak, and John Kominek, SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems, Interspeech 2007, Antwerp

SPICE 2008: Cross-continental Course

- SPICE-based course between CMU and UKA
  - Students at Carnegie Mellon University, PA
  - Students at Karlsruhe University, Germany
  - Linked by weekly meeting over VC
- Similar to 2007 BUT distributed collaboration
  - Students create ASR & TTS in their native language
  - Bonus for the ambitious: train SMT systems and create a speech-to-speech translation system
- Evaluation includes
  - Time to complete
  - Task difficulties
  - ASR word error rate
  - TTS voice quality

John Kominek, Sameer Badaskar, Tanja Schultz, Alan W Black, IMPROVING SPEECH SYSTEMS BUILT FROM VERY LITTLE DATA, submitted

Tutorial Outline

- Introduction and Motivation
  - Motivation
  - History and Leveraged Work
  - Rapid Language Adaptation Server: Spice
- SPICE in detail
  - Text collection & Prompt Selection
  - Phone set specification, Lexical construction
  - ASR Bootstrap & training
  - Language model, TTS Voice building
  - Testing and Tuning
- Latest Experiments and Results
  - Lessons Learnt from past studies
- Conclusions & Next Steps

Conclusions

- **Challenges in Multilingual Speech Processing**
  - Well defined build processes: ASR, MT, TTS … BUT:
  - Every new language brings unseen challenges
  - Current (statistical) approaches require lots of data
  - … and native language expert and technology expertise
  - How to bridge the gap between language and tech expert?
- **Proposed solution: SPICE**
  - Learning by interaction from a cooperative (but naïve) user
  - Rapid adaptation from language universal models
  - Knowledge sharing across components
  - Development cycle: Days rather than weeks

Next Steps

- **Continuous Server Support**
  - Improve Interface based on user feedback and lessons learned
  - Improve Language Robustness: font encoding, …
  - Software Engineering, Scaling
- **Collaboration**
  - Multiple people working on the same project
  - Leverage from archived projects
- **Cross-confirmation**
  - Multiple views for within and across project confirmation
  - Confidence measure to find appropriate combination
- **Error-blaming**
  - End-to-end system Evaluation vs Component Evaluation
  - Automatic Generation of Recommendations to improve systems

Try This At Home

- System is online at http://cmuspice.org
- Use system for your own project
  - Create new login/passwd and project
- Preloaded Hindi Example
  - Login as
    - Login: demo
    - Passwd: demo
  - Chose project # (your birth day)
- Book on ML Speech Processing
  Elsevier, Academic Press, 2006