Rapid Language Portability of Speech Processing Systems

Tanja Schultz

Language Technologies Institute, InterACT,
Carnegie Mellon University
MULTILING, Stellenbosch, April 10, 2006

Motivation

❶ Computerization: Speech is key technology

➡ Mobile Devices, Ubiquitous Information Access

❷ Globalization: Multilinguality

➡ More than 6900 Languages in the world

➡ Multiple official languages

➡ Europe has 20+ official languages

➡ South Africa has 11 official languages

⇒ **Speech Processing in multiple Languages**

➡ Cross-cultural Human-Human Interaction

➡ Human-Machine Interface in mother tongue

Challenges

- Algorithms are language independent but require data
  - → Dozens of hours audio recordings and corresponding transcriptions
  - → Pronunciation dictionaries for large vocabularies (>100.000 words)
  - → Millions of words written text corpora in various domains in question
  - → Bilingual aligned text corpora
- BUT: Such data are only available in very few languages
  - → Audio data ≤ 40 languages, Transcriptions take up to 40x real time
  - → Large vocabulary pronunciation dictionaries ≤ 20 languages
  - → Small text corpora ≤ 100 languages, large corpora ≤ 30 languages
  - → Bilingual corpora in very few language pairs, pivot mostly English
- Additional complications:
  - → Combinatorical explosion (domain, speaking style, accent, dialect, ...)
  - → Few native speakers at hand for minority (endangered) languages
  - → Languages without writing systems

Solution: Learning Systems

⇒ Intelligent systems that learn a language from the user

- Effizient 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): by a factor of 10
    - → Speed up development cycles: days rather than months

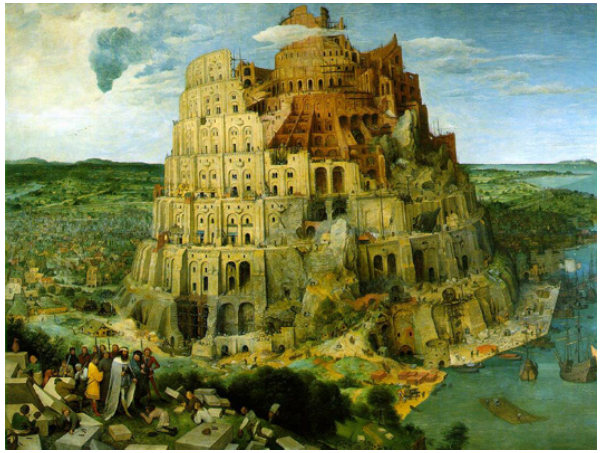
  ⇒ Rapid Language Adaptation from universal models

- 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

SPICE

Speech **P**rocessing: **I**nteractive **C**reation and **E**valuation toolkit

- National Science Foundation, Grant 10/2004, 3 years
- Principle Investigators Tanja Schultz and 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*

SPICE

Speech Processing - Interactive Creation and Evaluation Toolkit for New Languages

Phone Selection | **Acoustic Model** | **Language Model** | **Dictionary** | **TTS**

Welcome to SPICE homepage. In this page, you can develop the speech system (including Automatic Speech Recognition and Text-To-Speech) specific to your own language.

First thing to your attention:

Please ENABLE YOUR BROWSER TO ACCEPT COOKIES because in SPICE web site, we use cookies to identify different users and different projects. If your browser has not been enabled to accept cookies, then you cannot login and use SPICE at all.

Please input your name (Developer Name) and the name of your language (Language Name) and the name of the Project (Project Name) below. The basic idea is that each developer can work on several different languages, and in each language you can develop several different version of your speech system, differentiated by "Project Name".

You don't have to finish all the work at one time. For example, you can do Acoustic Model at one day, Language Model tomorrow ...etc. Every time you come back to SPICE page, Use your "Developer Name" , "Language Name" and "Project Name", you can locate your previous work and continue to do the rest.

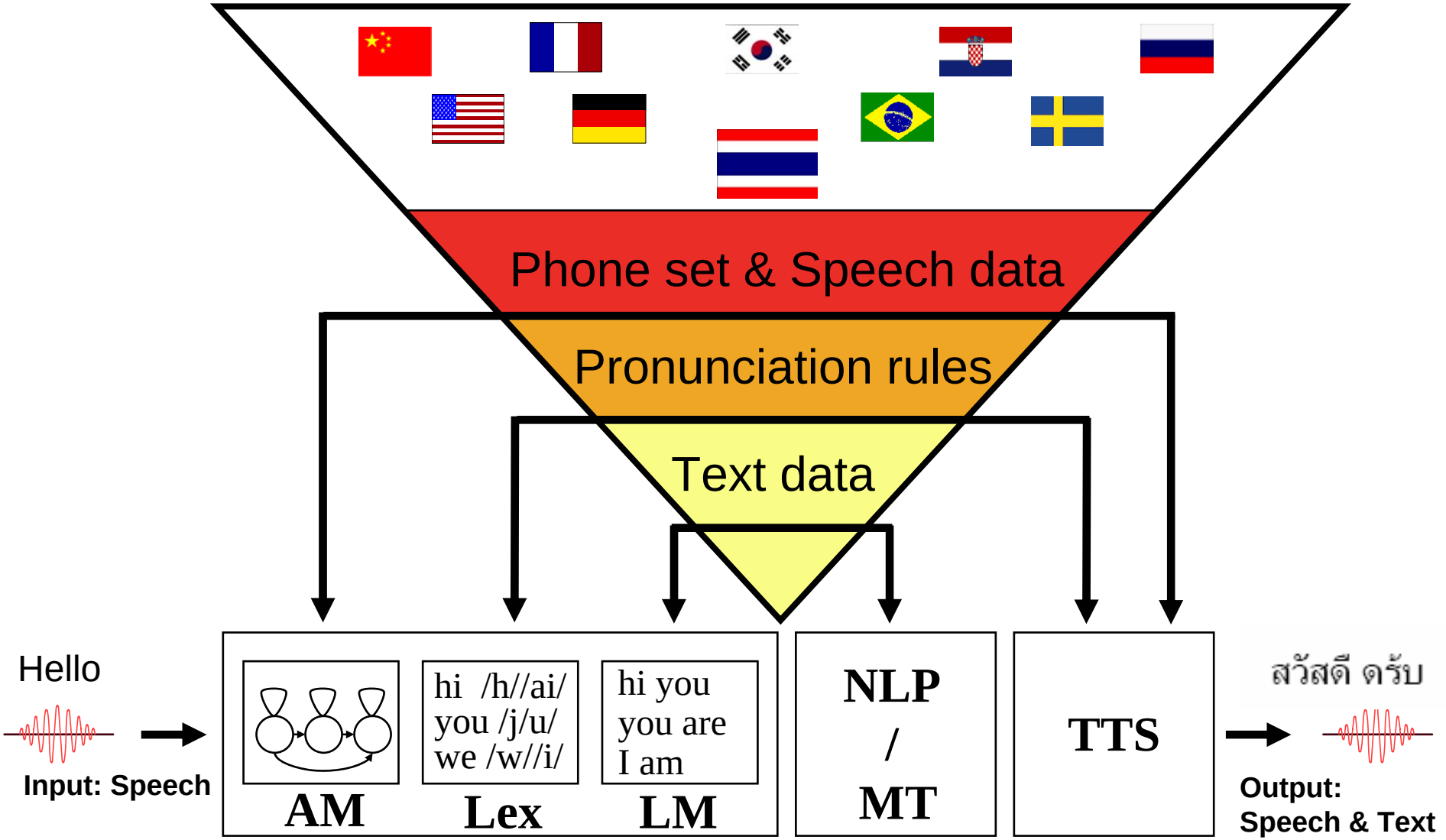
Developer Name (No space please):

Language Name (No space please):

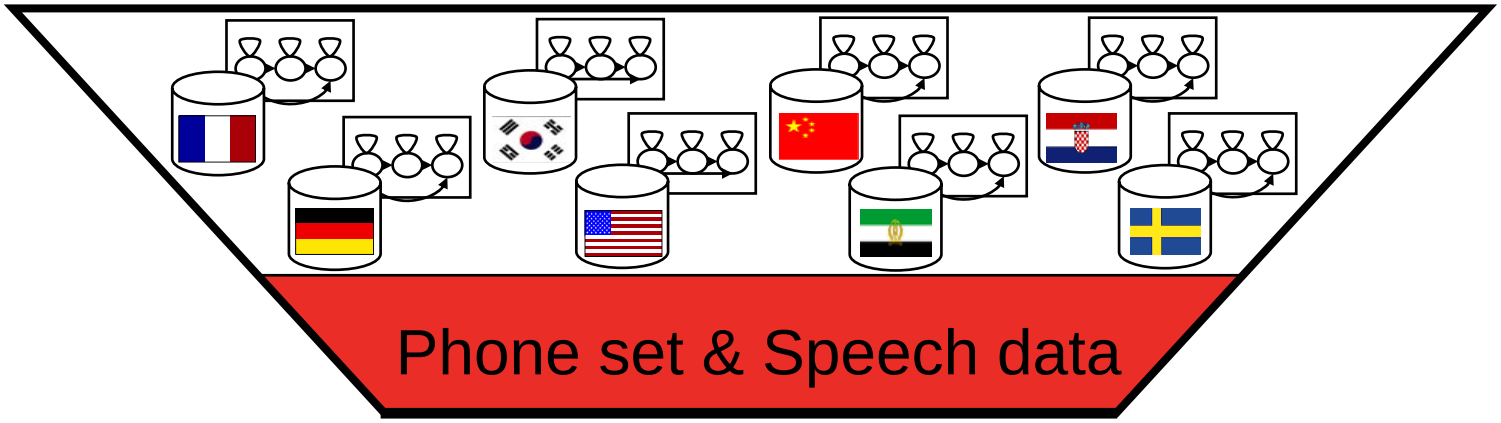
Project Name (No space please):

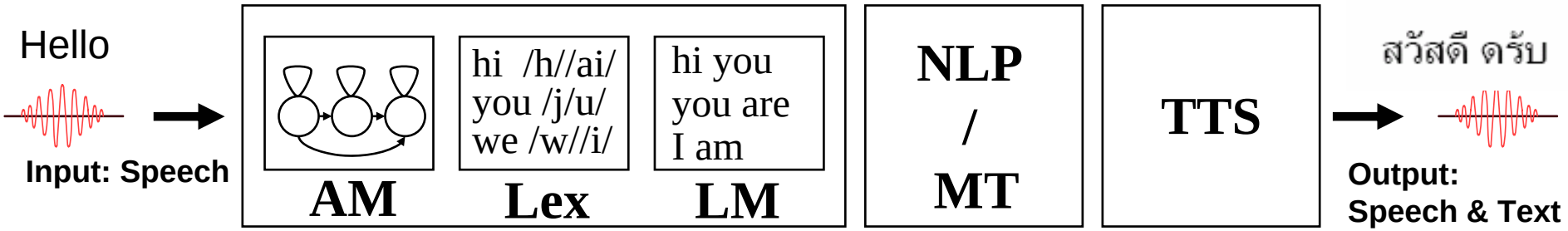
Submit

Speech Processing Systems



Rapid Portability: Data





GlobalPhone

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

Multilingual Database

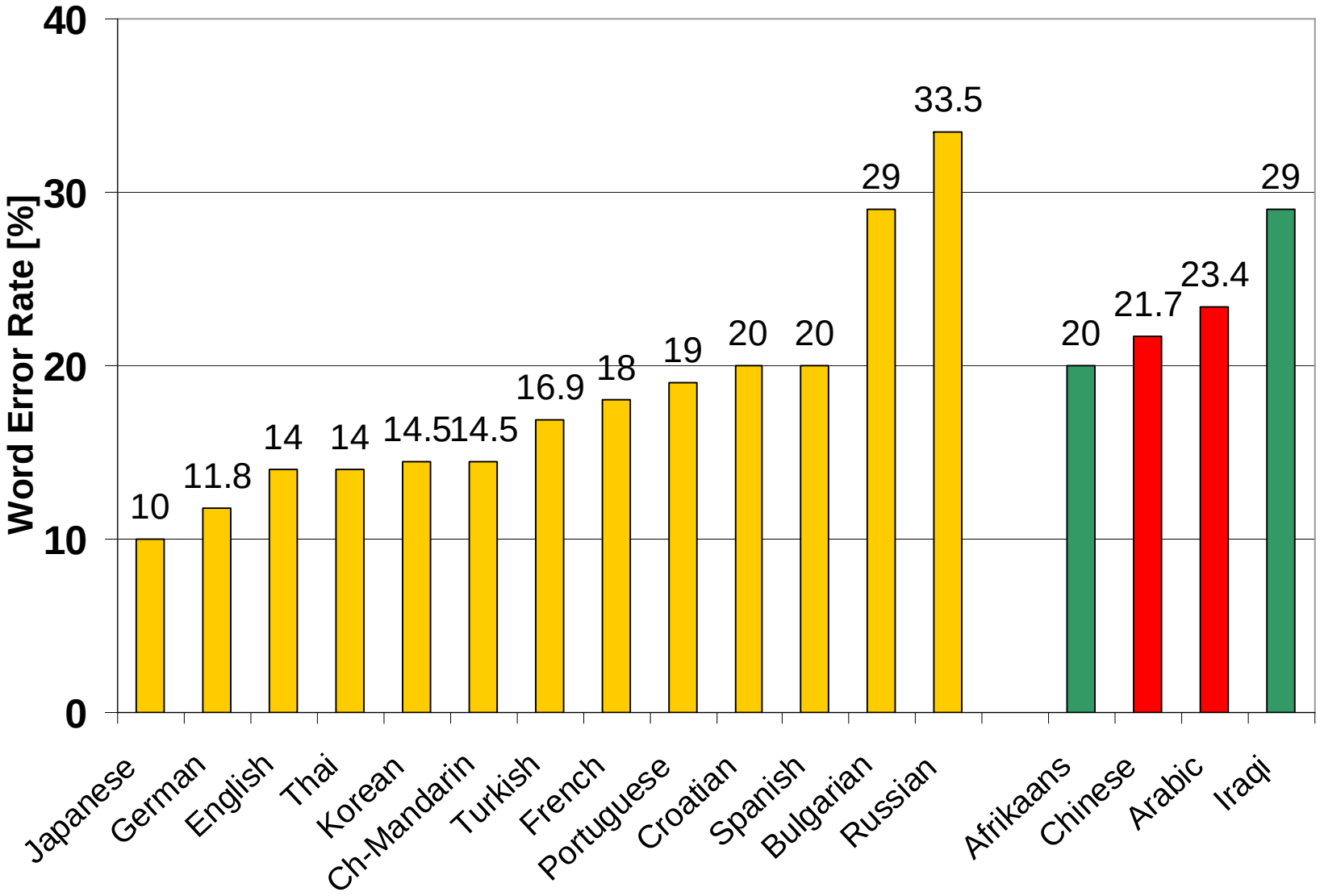
- Widespread languages
- Native Speakers
- Uniform Data
- Broad Domain
- Large Text Resources
  - ➡ Internet, Newspaper

Corpus

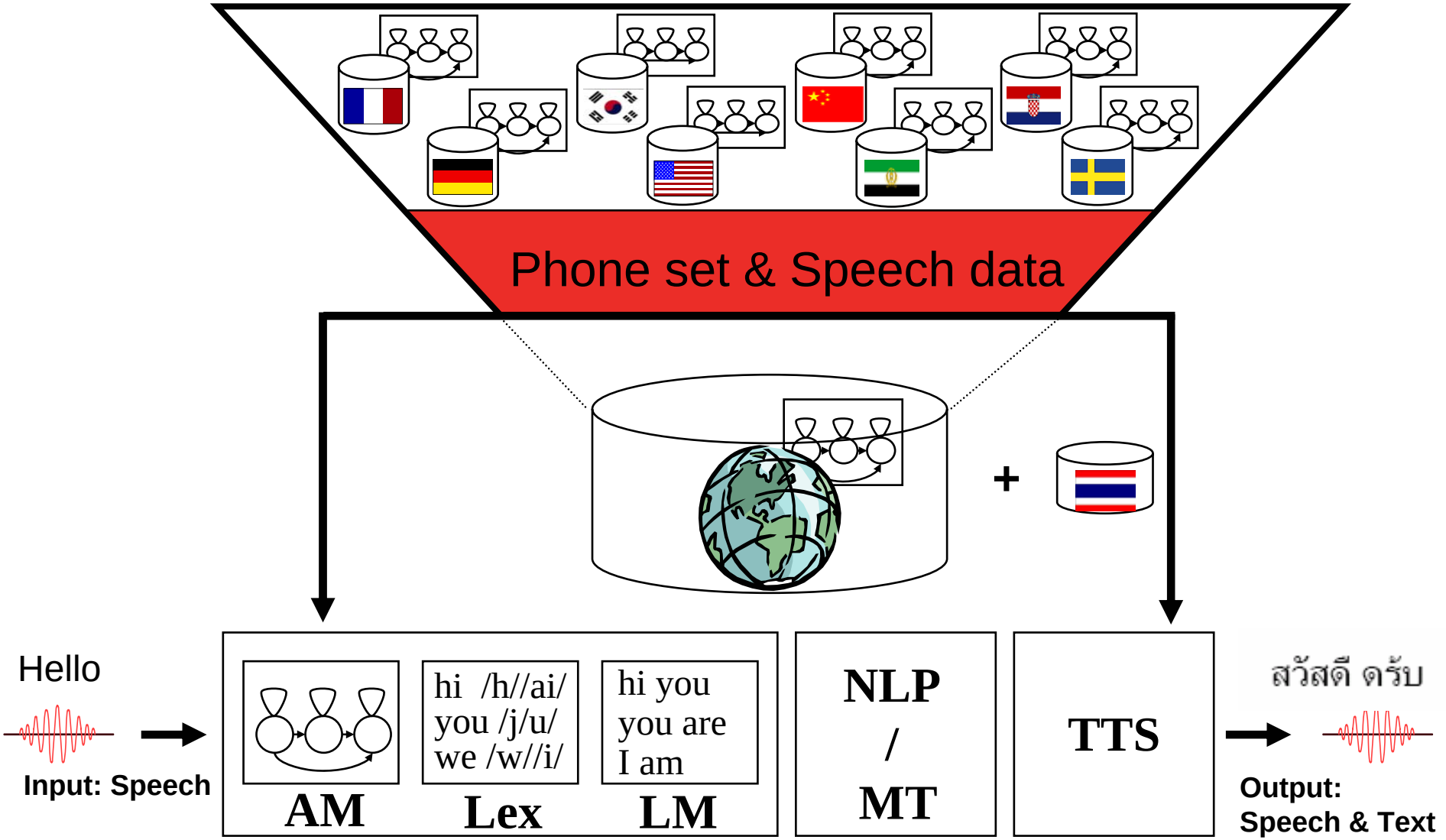
- 19 Languages … counting
- ≥ 1800 native speakers
- ≥ 400 hrs Audio data
- Read Speech
- Filled pauses annotated

Now available from ELRA !!

Speech Recognition in 17 Languages



Rapid Portability: Acoustic Models



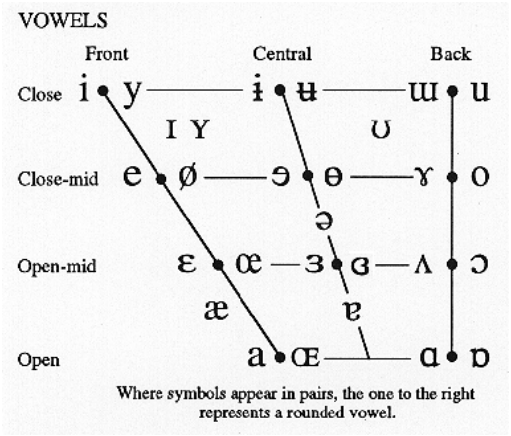
Universal Sound Inventory

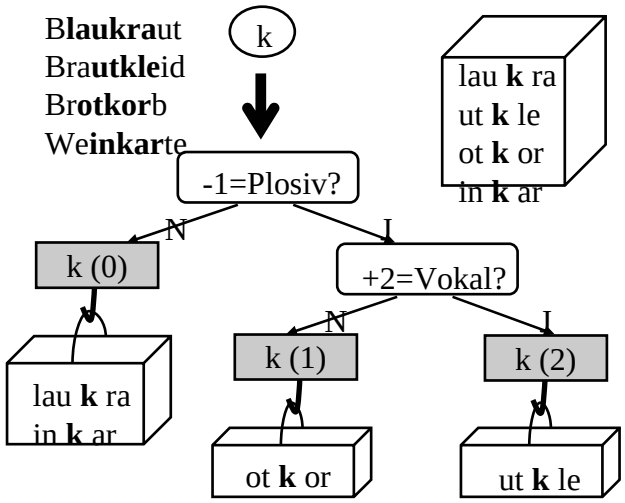
Speech Production is independent from Language ⇒ IPA

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





Problem:

Context of sounds are language specific

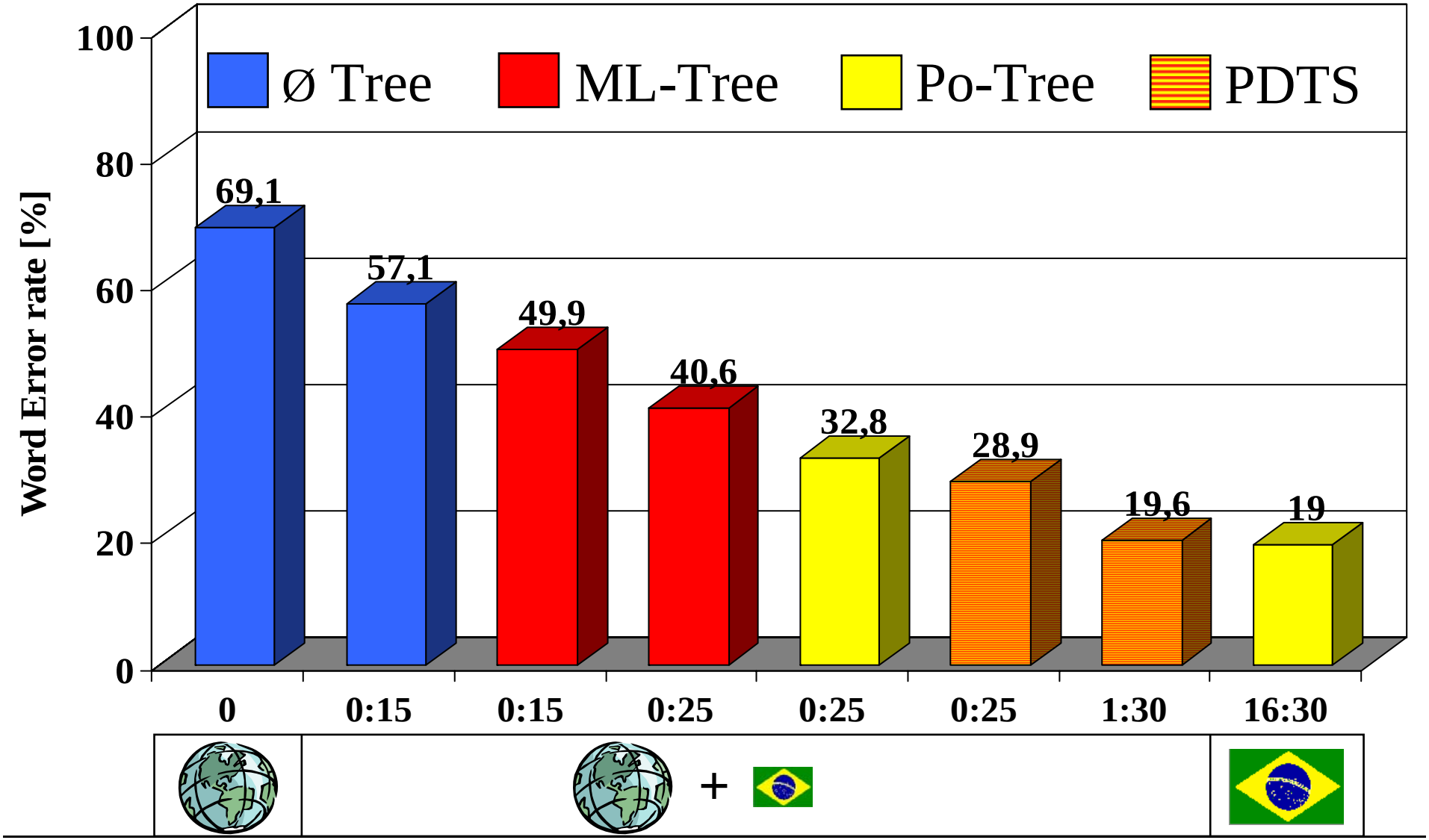
Context dependent models for new languages?

Solution:

1) Multilingual Decision Context Trees

2) Specialize decision tree by Adaptation

Rapid Portability: Acoustic Model



SPICE

Speech Processing - Interactive Creation and Evaluation Toolkit for New Languages

Acoustic Model | **Language Model** | **Dictionary** | **T T S**

Switch to page without example sound

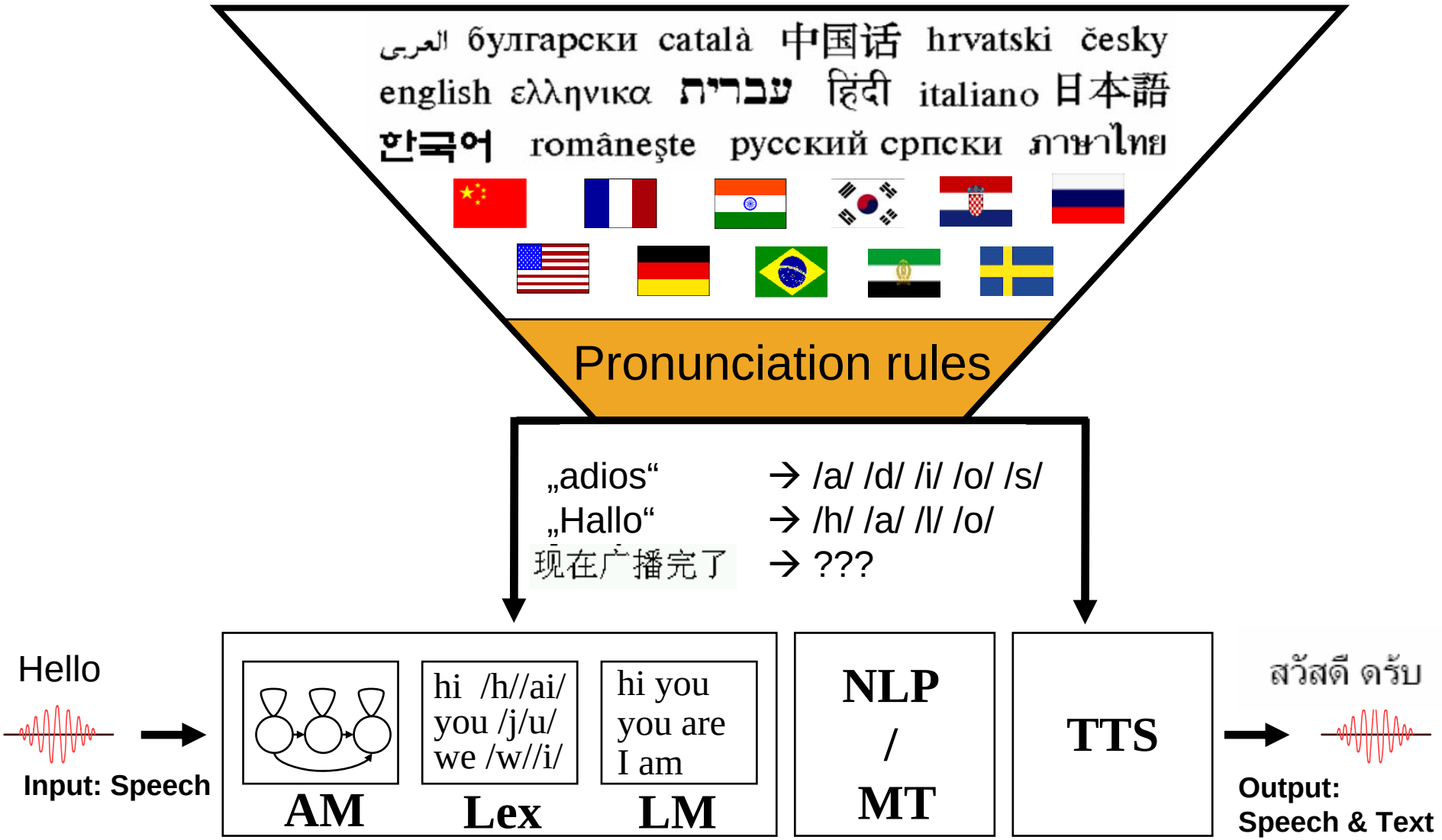
Currently you are logging in as tanja, you are working on language German , Project Name = 1

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

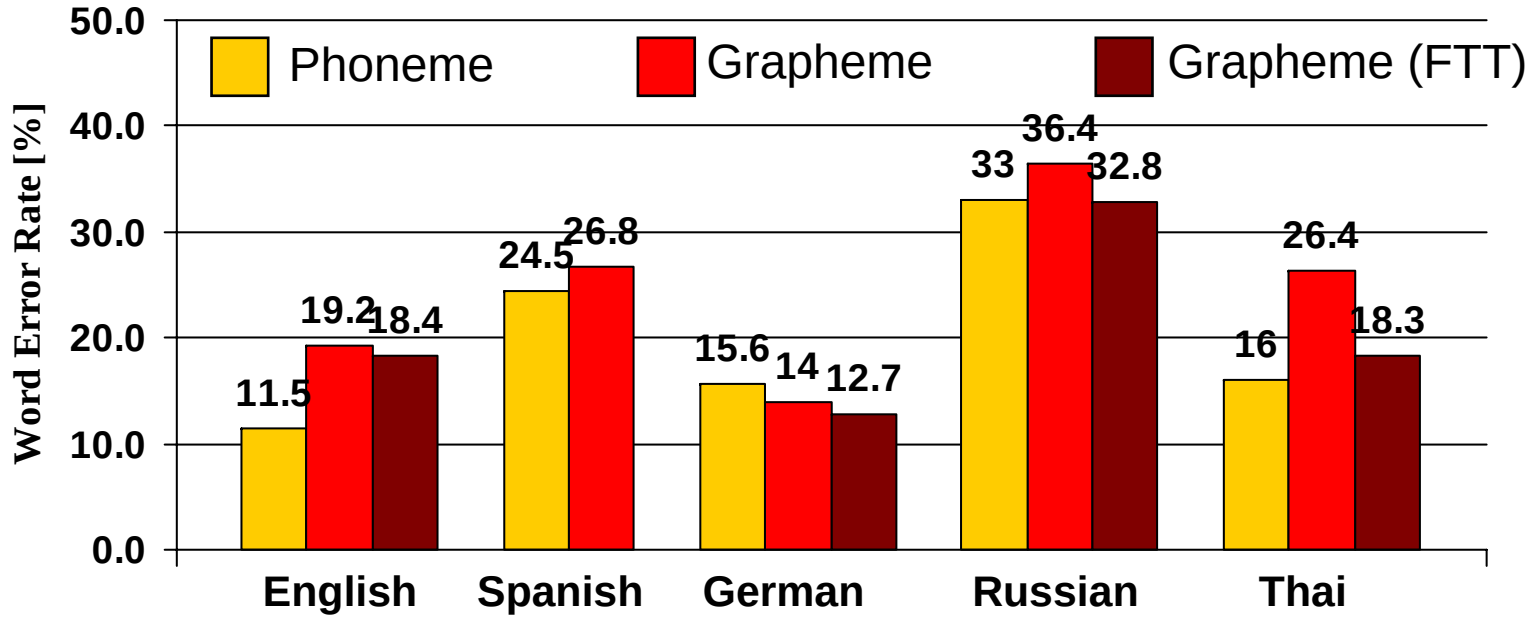
Consonants (Pulmonic): Please choose the consonant sounds you'd like to have in your new acoustic models by giving it a name in the textbox next to it.

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p pʲ b bʲ			t tʲ d dʲ		ʈ ɖ	c ɟ	k kʲ g gʲ	q ɢ		ʔ
Nasal	m mʲ	m		n nʲ		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f fʲ v vʲ	θ ð	s sʲ z zʲ	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							

Rapid Portability: Pronunciation Dictionary



Phoneme- vs Grapheme based ASR



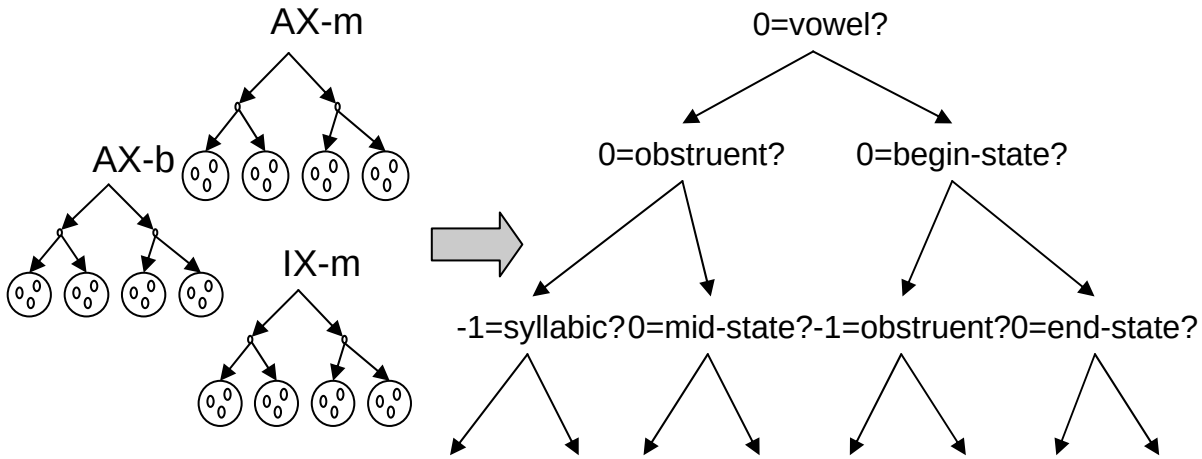
Problem:

- 1 Grapheme ≠ 1 Phoneme

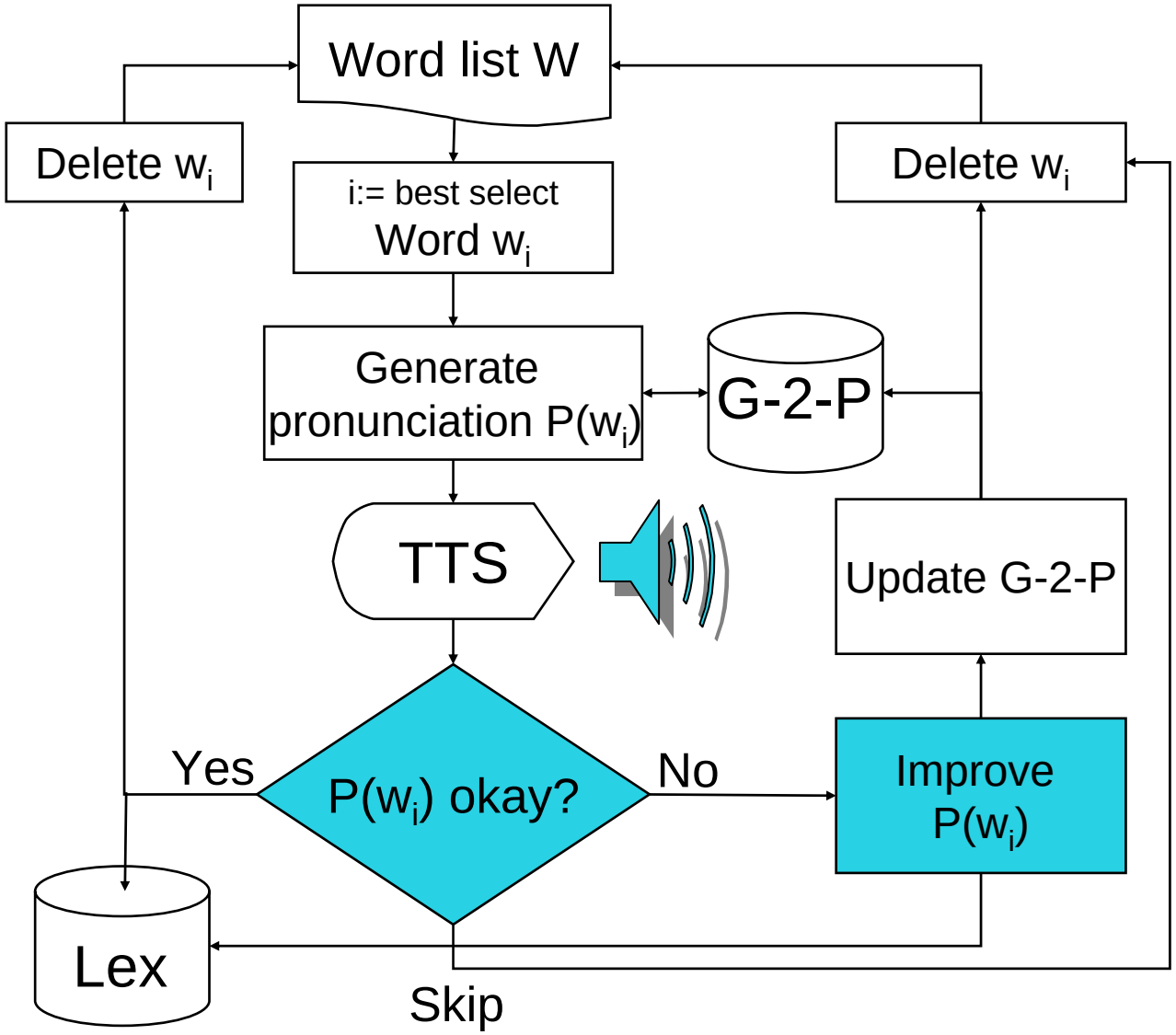
Flexible Tree Tying (FTT):

One decision tree

- Improved parameter tying
- Less over specification
- Fewer inconsistencies



Dictionary: Interactive Learning



* Follow the work of Davel&Barnard

* Word list:
  extract from text

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

* Update after each w_i
→ more effective training

User

SPICE

Speech Processing - Interactive Creation and Evaluation Toolkit for New Languages

Phone Selection | **Acoustic Model** | **Language Model** | **Dictionary** | **TTS**

Currently you are logging in as tanja, you are working on language German , Project Name = 1

We use an iterative procedure to gather information to create dictionary for you.

First of all, please input an initial Grapheme to Phoneme (G2P) rule of your language.

Based on this rule, our system will "guess" the correct pronunciation of 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 our system 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 type one of the most common pronunciation for each grapheme. Thanks.

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

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

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

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

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

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

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

11.35593220339% Finished
new word:

Merck

system suggested pronunciation: m e r x k 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

Here's a list of all phonemes you selected in AM page:

p
b
t
d
k
g
m
n
r
f
v
s
z
x
i
u
e
o
a

Issues and Challenges

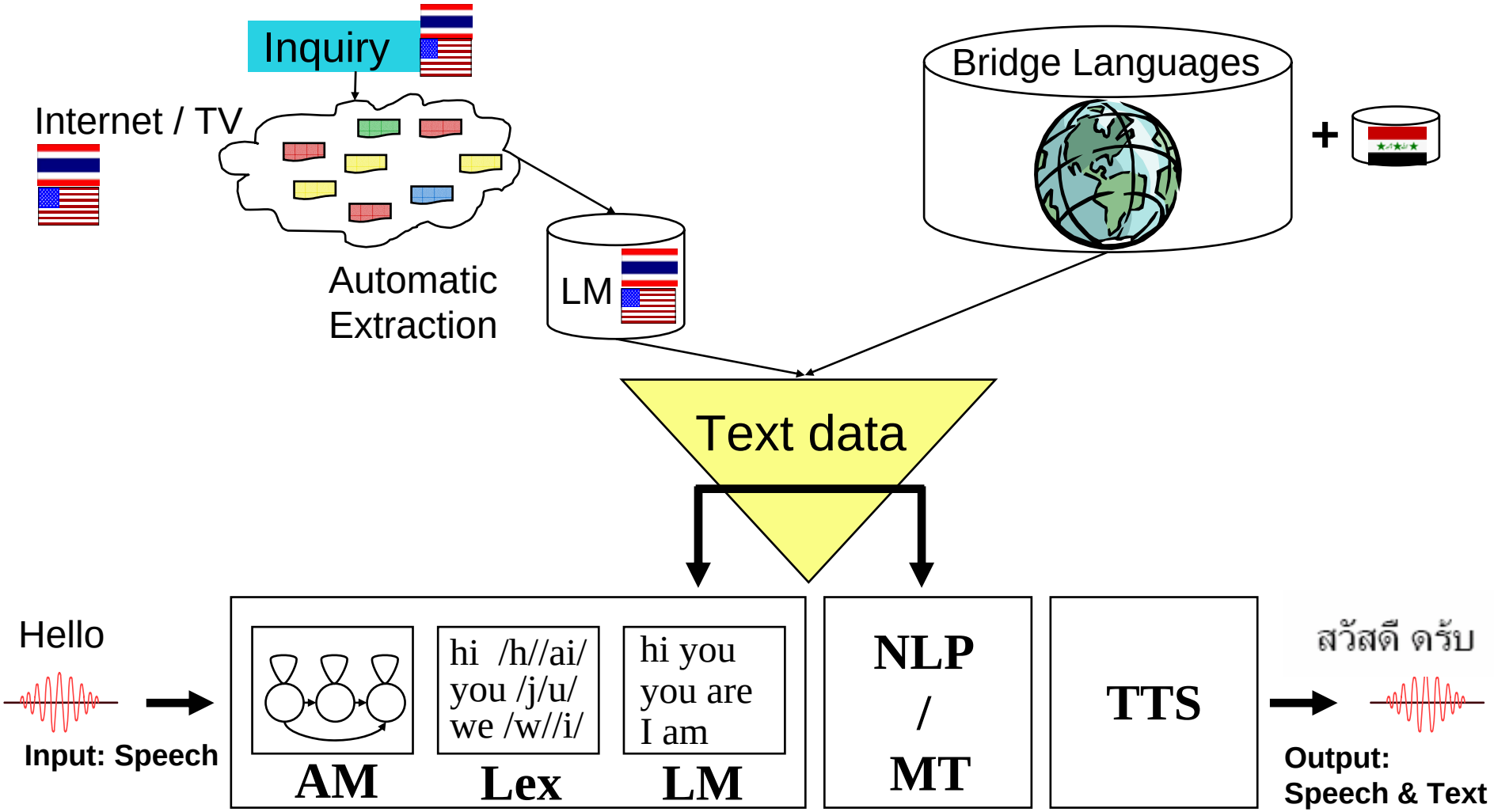
- How to make best use of the human?
  - Definition of successful completion
  - Which words to present in what order
  - How to be robust against mistakes
  - Feedback that keeps users motivated to continue

- How many words to be solicited?
  - G2P complexity depends on language
  - 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

Rapid Portability: LM

Resource rich languages ↔ Resource low languages:



SPICE

Speech Processing - Interactive Creation and Evaluation Toolkit for New Languages

Phone Selection | Acoustic Model | Language Model | Dictionary | TTS

Build your own language model

Currently you are logging in as tanja, you are working on language e , Project Name = 1

We will try to grab text data from internet and build language model from those text data.

Given a link to some webpage composed of your own language, our web spider can automatically go into this web page, grab all text from this page, and also follow the hyperlink inside this page then go to another linked page to collect more text data. We need huge text data in order to build a robust language model. As we often said in speech research community, "no data is like more data".

Please provide us a URL of webpage which is composed by your language:

Please specify how many levels would you like our web spider to get into:

(•) 1 ( ) 2

Please verify that the URL you gave above is correct and valid, select the level you want the web spider to get. The deeper the value is, the longer it takes to finish the task. So our suggestion is to select a smaller value unless you care about the amount of training text more than the time to finish the task.

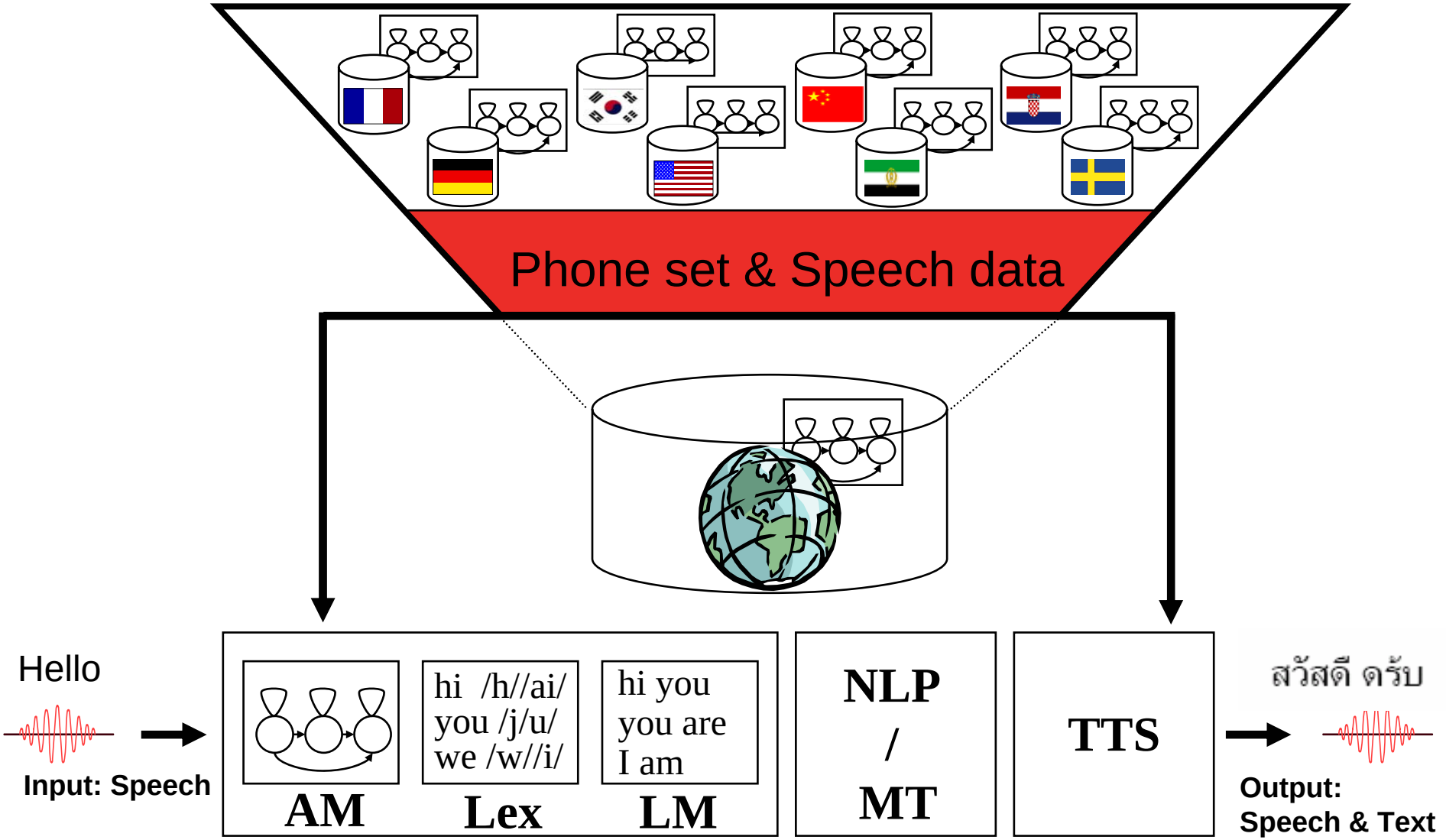
Another way is to submit a text file from your local machine. This file should contains large amount of text of your language. We can train Language Model from this text as well. But if you want to submit a text file, please make it such that every line contains a single sentence. This is the standard input format our LM toolkit will take.

You can also train LM via both method. Just input both the URL and the file name. Our toolkit will use both information.

Choose a file to upload: Browse...

Submit

Rapid Portability: TTS

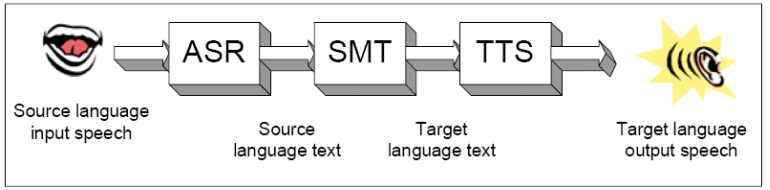


Parametric TTS

- Text-to-speech for G2P Learning:
  - Technique: phoneme-by-phoneme concatenation, speech not natural but understandable (Marelie Davel)
  - Units are based on IPA phoneme examples
    - → PRO: covers languages through simple adaptation
    - → CONS: not good enough for speech applications
- Text-to-speech for Applications:
  - Common technologies
    - → Diphone: too hard to record and label
    - → Unit selection: too much to record and label
  - New technology: ***clustergen*** trajectory synthesis
    - → Clusters representing context-dependent allophones
    - → PRO: can work with little speech (10 minutes)
    - → CONS: speech sounds buzzy, lacks natural prosody

SPICE: Afrikaans - English

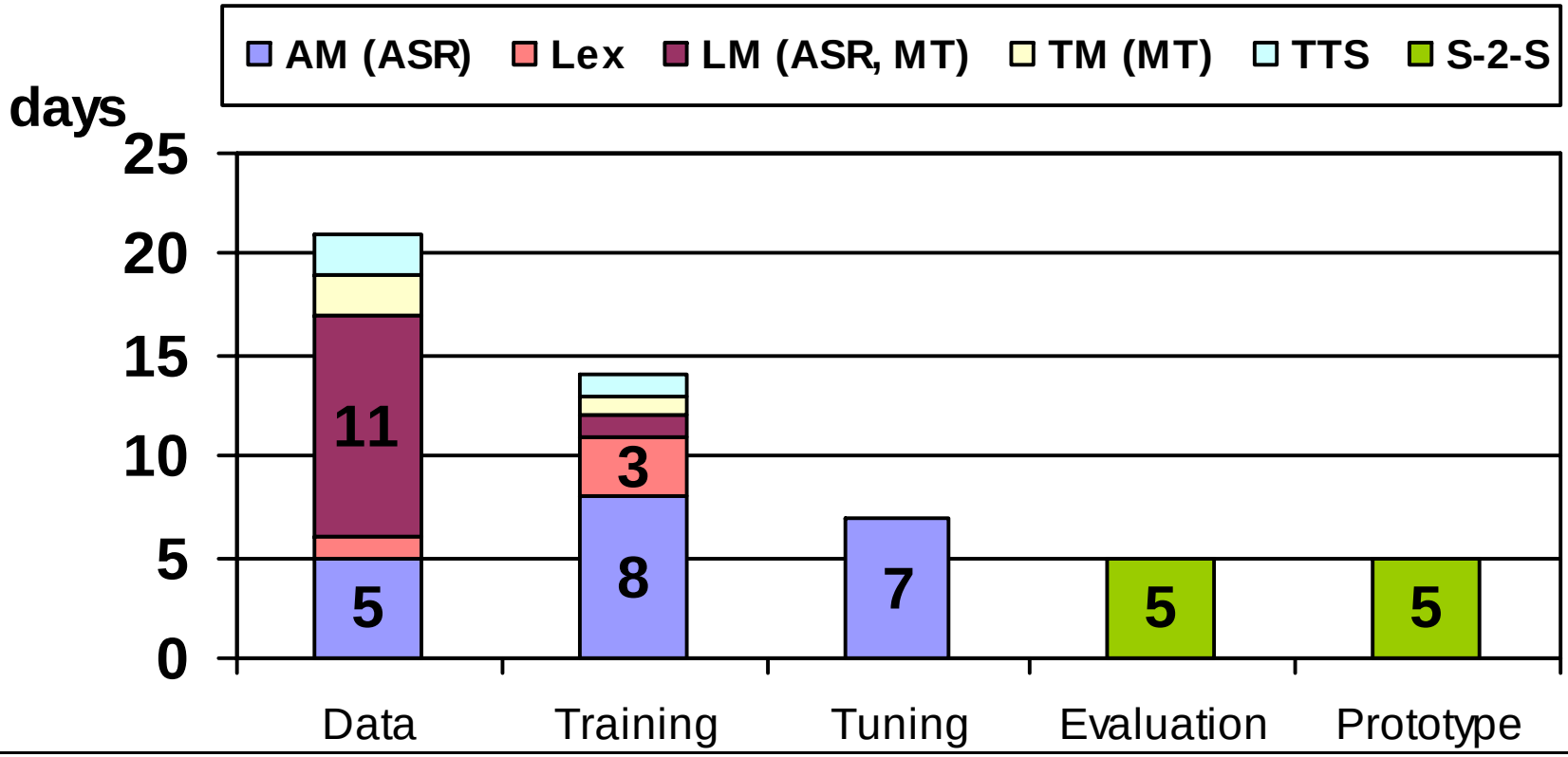
- Goal: Build Afrikaans – English Speech Translation System using SPICE
  - → Cooperation with University Stellenbosch and ARMSCOR
  - → Bilingual PhD visited CMU for 3 month (thanks Herman Engelbrecht !!!)
  - → Afrikaans: Related to Dutch and English,
    g-2-p very close, regular grammar, simple morphology
- 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)

SPICE: 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!)



Other Projects on Multilinguality

- Constantly growing interest in multilinguality
- Major needs:
  - Information gathering from multiple sources
  - Translation requirements for multilingual communities
  - Two-way communication
- Translation of BN, Lectures, and Meetings
  - US: GALE (DARPA), STR-Dust (NSF)
  - Europe: TC_Star (EU FP6)
- Translation in mobile communication scenarios
  - US: TransTac (DARPA), Thai ST (Laser)

Translation of Broadcast News, Lectures and Meetings

Projects:
- TC_STAR (EC FP6)
- STR-DUST (NSF)
- Gale (DARPA)

Demo

Gale: Global Autonomous Language Exploitation

- Largest DARPA project in HLT (EARS+TIDES)
- Automatically process huge volumes of speech and text data in *multiple languages*
  - Broadcast News, Talk Shows, Telephone Conversations
  - Chinese, Arabic (+ dialectal variations), surprise languages
- Deliver pertinent information in easy-to-understand forms to monolingual analysts, 3 engines:
  - Transcription: Transform multilingual speech to text
  - Translation: transform any text to English
  - Distillation: extract & present information to English analyst

Demonstration

Mandarin
Broadcast News
CCTV
recorded in the US
over satellite

Transforming the Mandarin speech Into Chinese text using **Automatic Speech Recognition**

Translating from Chinese text into English text using **Statistical Machine Translation**

PDA Speech Translation in Mobile Scenarios

- Tourism
  - Needs in Foreign Country
  - International Events
    - → Conferences
    - → Business
    - → Olympics
- Humanitarian Needs
  - Humanitarian, Government
    - → Medical, Refugee Registration
- Projects:
  - Thai ST (Laser)
  - TransTac (DARPA)

TransTac

- Team effort:
  - Speech Recognition (CMU / Mobile, LLC)
  - Statistical MT (CMU / Mobile, LLC)
  - Speech Synthesis Swift (Cepstral, LLC)
  - Graphical User Interface (Mobile, LLC)
- System runs on all platforms
  - Off-the-shelf consumer PDAs
  - Laptop/Desktop under Win/CE/Linux
  - Phraselator P2 (Voxtec)
- Interface
  - Simple and intuitive push-to-talk
  - Back translation for confirmation

- Language pairs: English-Thai + English-Arabic
- Handheld: Joint optimization of speed and accuracy
  - About 1.5 real-time on a 800MHz PXA270, 128Mb RAM

Conclusion

- **Intelligent systems to learn language**
  - SPICE: Learning by interaction with the (naive) user
  - Rapid Portability to unseen languages
- **Multilingual Systems**
  - Systems and data in multiple languages
  - Universal language independent models
- **Projects on Multilinguality**
  - Extract information from multilingual speech data
  - Speech translation in mobile scenarios

Products

Multilingual Speech Processing
Book information
Product description
Audience
Author information and services
Ordering information
Bibliographic and ordering information
Conditions of sale
Book related information
Submit your book proposal
Other books in same subject area

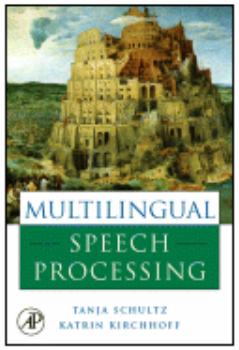
Support & contact

About Elsevier

Select your view

MULTILINGUAL SPEECH PROCESSING

To order this title, and for more information, click here
First Edition



Edited By
Tanja Schultz, Professor, Carnegie Mellon University, Pittsburgh, PA, USA
Katrin Kirchhoff, Professor, University of Washington, Seattle, WA, USA

Description

Tanja Schultz and Katrin Kirchhoff have compiled a comprehensive overview of speech processing from a multilingual perspective. By taking this all-inclusive approach to speech processing, the editors have included theories, algorithms, and techniques that are required to support spoken input and output in a large variety of languages. This book presents a comprehensive introduction to research problems and solutions, both from a theoretical as well as a practical perspective, and highlights technology that incorporates the increasing necessity for multilingual applications in our global community. Current challenges of speech processing and the feasibility of sharing data and system components across different languages guide contributors in their discussions of trends, prognoses and open research issues. This includes automatic speech recognition and speech synthesis, but also speech-to-speech translation, dialog systems, automatic language identification, and handling non-native speech. The book is complemented by an overview of multilingual resources, important research trends, and actual speech processing systems that are being deployed in multilingual human-human and human-machine interfaces. Researchers and developers in industry and academia with different backgrounds but a common interest in multilingual speech processing will find an excellent overview of research problems and solutions detailed from theoretical and practical perspectives.

Audience

Researchers & government employees in industry, consultants in speech/signal processing, undergraduate and graduate students

Contents

CH 1: Introduction CH 2: Language Characteristics CH 3: Linguistic Data Resources CH 4: Multilingual Acoustic Modeling CH 5: Multilingual Dictionaries CH 6: Multilingual Language Modeling CH 7: Multilingual Speech Synthesis CH 8: Automatic Language Identification CH 9: Other Challenges CH 10: Speech-to-Speech Translation CH 11: Multilingual Spoken Dialog Systems Bibliography

Bibliographic & ordering Information

Hardbound, ISBN: 0-12-088501-8, 480 pages, publication date: 2006
Imprint: ACADEMIC PRESS
Price: Order form
USD 69.95
GBP 39.99
EUR 57.95

Book contents
- Table of contents

Reviews
- Submit your review

Free e-mail alerting services
For book tables-of-contents of forthcoming Elsevier books
CONTENTS direct

Bookmark this page
Recommend this publication
Overview of all books

Interspeech2006 ICSLP

International Conference on Spoken Language Processing

17-21 September 2006 Pittsburgh PA, USA

Home
Call for Papers
Special Sessions

About Pittsburgh
Venue

Archives

Welcome to INTERSPEECH 2006 — ICSLP

The Ninth International Conference on Spoken Language Processing (Interspeech 2006 — ICSLP) will be held in **Pittsburgh, Pennsylvania**, under the sponsorship of the International Speech Communication Association (ISCA). Interspeech 2006 — ICSLP, follows on from Interspeech 2004 — ICSLP, in Jeju, Korea, October 2004 and Interspeech 2005 — Eurospeech Lisbon, Portugal, September 2005. Today INTERSPEECH, the continuation of the ICSLP and Eurospeech conferences, enjoys ever-increasing impact and influence as the focal point for the exchange of ideas in a broad array of fields centered around human-human and human-machine speech communication.

Interspeech 2006 will cover all aspects of speech science and technology. The conference will include plenary talks by world-class experts, tutorials, exhibits, special sessions covering interdisciplinary topics and important new emerging areas of interest, and parallel oral and poster sessions. **Topic areas** are listed on the Call for Papers webpage.

The **conference venue** will be the The Westin Convention Center Pittsburgh.
Note: This is the hotel next to the Pittsburgh Convention Center.

Special Sessions will include:

- **The Speech Separation Challenge**. Martin Cooke, Sheffield, and Te-Won Lee, UCSD.
- **Speech Summarization**. Jean Carletta, Edinburgh, and Julia Hirschberg, Columbia.
- **Articulatory Modeling**. Eric Bateson, British Columbia.
- **Visual Intonation**. Marc Swerts, Tilburg.
- **Spoken Dialog Technology R&D**. Roberto Pieraccini, TellEureka.
- **The Prosody of Turn-Taking and Dialog Acts**. Nigel Ward, UT El Paso, and Elizabeth Shriberg, SRI and ICSI.
- **Speech and Language in Education**. Patti Price, pprice.com, and Abeer Alwan, UCLA.
- **From Ideas to Companies**. Janet Baker, formerly of Dragon.