Multilinguality
SPICE: making it easier
Dealing with *all* Languages

- **Over 6000 Languages**
  - Maybe not all commercially interesting … now

- **Major languages (economic)**
  - Cell phone manufacturers list 46 languages
  - But even those not all covered
Motivation

Computerization: Speech is key technology

Mobile Devices, Ubiquitous Information Access

Globalization: Multilinguality

More than 6000 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 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 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

⇒ 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): by a factor of 10
    ▶ Speed up development cycles: days rather than months

⇒ Rapid Language Adaptation from universal models

◆ Bridge the gap: language and technology experts

  ▶ Technology experts do not speak all languages in question
  ▶ Native users are not in control of the technology
Sharing data between modules

Speech-to-Speech Translation

Input $L_s$

Output $L_s$

$L_{source}$

$L_{target}$
**Speech Processing: Interactive Creation and Evaluation 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](http://cmuspice.org)
CMU SPICE

Build Your System

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

User: awb Language: eng Project: aug10 [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
Speech Processing Systems

Phone set & Speech data

Pronunciation rules

Text data

Hello

AM
Lex
LM
NLP
MT
TTS

Input: Speech

Output: Speech & Text

AM
Lex
LM
NLP
MT
TTS

Input: Speech

Output: Speech & Text
Rapid Portability: Data

Phone set & Speech data

Input: Speech

AM
Lex
LM

NLP / MT

TTS

Output: Speech & Text

Hello

/r/ /h//ai/
you /j/u/
we /w//i/
hi you
you are
I am
Finding “Nice” Prompts

- From very large text databases
- Find “nice” sentences:
  - Containing only high frequency words
  - 5-15 words
- Find grapheme/phoneme balanced set
  - Select sentences with best triphone/graph
- 500-1000 sentences
- Collect for ASR and TTS acoustic modeling
Prompt Selection Issues

- **Need good text**
  - De-htmlify, well-written, no misspelling

- **Need word segmentation**
  - Japanese, Chinese Thai

- **Natural text is often mixed language**
  - Hindi Newspaper Text has lots of English words

- **Automatic selection has errors**
  - Need Speaker to do further selection
  - E.g. lots of telephone numbers, formating commands

- **CMU Arctic used similar methods**
Recording Prompts

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]

Current Status

- Speaker ID: 006
- Speaker Name: Tanja Schultz
- Prompts Completed: of 4 (NOT-RECORDED)

File: prompt_001.wav  Length: 0.0  Position: 0.0

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.ic.cmu.edu:7800
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

Word Error Rate [%]

Japanese 10 11.8 14 14 14.5 14.5
German 16.9 18 19 20 20
English Russian
Turkish 29
French Afrikaans
Portuguese Chinese
Croatian Arabic
Spanish 33.5
Bulgarian Iraqi
Russian
Rapid Portability: Acoustic Models

Phone set & Speech data

Input: Speech

Hello

Output: Speech & Text

AM
Lex
LM
NLP
MT
TTS
Speech Production is independent from Language $\Rightarrow$ 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
Choosing Phonemes


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.
Rapid Portability: Acoustic Model

Word Error rate [%]

- Ø Tree
- ML-Tree
- Po-Tree
- PDTS

<table>
<thead>
<tr>
<th>Time</th>
<th>Ø Tree</th>
<th>ML-Tree</th>
<th>Po-Tree</th>
<th>PDTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:00</td>
<td>69.1</td>
<td>57.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0:15</td>
<td>49.9</td>
<td>40.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0:25</td>
<td>32.8</td>
<td>28.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:30</td>
<td>19.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td>19.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rapid Portability: Pronunciation Dictionary

Pronunciation rules

"adios" → /a/ /d/ /i/ /o/ /s/
"Hallo" → /h/ /a/ /l/ /o/ → ???

Hello

Input: Speech

AM

Lex

LM

NLP

MT

TTS

Output: Speech & Text

你好
Problem:
• 1 Grapheme ≠ 1 Phoneme

Flexible Tree Tying (FTT):
*One* decision tree
• Improved parameter tying
• Less over specification
• Fewer inconsistencies
* 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

* Kominek & Black
User: awb Language: eng Project: aug19

Lexicon pronunciation creation

Rule entry

3.0075187969925% Finished

new word:

at

system suggested pronunciation: [AX T]

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

Grapheme-to-phoneme rules (help)

Phoneme labels for your language:
P B T D K G M

Audio collection (help)

Text and prompt selection (help)
User: awb Language: eng Project: aug19  [Logout]

Lexicon pronunciation creation

Rule entry

3.5087719298246% Finished

new word: Jeanne

system suggested pronunciation: * AX N N

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
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?
- G2P complexity language dependent
- 80% coverage
  - hundred (SP) to thousands (EN)
- G2P rule system perplexity

<table>
<thead>
<tr>
<th>Language</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>50.11</td>
</tr>
<tr>
<td>Dutch</td>
<td>16.80</td>
</tr>
<tr>
<td>German</td>
<td>16.70</td>
</tr>
<tr>
<td>Afrikaans</td>
<td>11.48</td>
</tr>
<tr>
<td>Italian</td>
<td>3.52</td>
</tr>
<tr>
<td>Spanish</td>
<td>1.21</td>
</tr>
</tbody>
</table>
Rapid Portability: LM

Resource rich languages ↔ Resource low languages:

Input: Speech

AM
Lex
LM

NLP / MT

TTS

Output: Speech & Text

Internet / TV

Automatic Extraction

Inquiry

Bridge Languages

Resource rich languages

Resource low languages:

Text data

Hello

สวัสดี ครับ
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:
- Statistical Parametric Systems: clustergen
- Clusters representing context-dependent allophones
  - PRO: can work with little speech (10 minutes)
  - PRO: robust to erroneous data.
  - CONS: speech sounds buzzy, lacks natural prosody
Goal: Build Afrikaans – English S2S using SPICE
- Cooperation with University Stellenbosch and ARMSCOR
- Bilingual PhD visited CMU for 3 month (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 hansard; 680k words;
- 43k bilingual aligned sentence pairs;

Audio: 6 hours read speech; 10k utterances,
SPICE: Time effort

- 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!)

![Bar chart showing time effort in days for AM (ASR), Lex, LM (ASR, MT), TM (MT), TTS, and S-2-S.]
Current Tests

- 11 students is CMU class
  
- Hindi (2), Vietnamese (2), French, German (2), Bulgarian, Telugu, Cantonese, Mandarin.

- Build complete S2S system
  
  - Teams of 2 for translation on small domain
  
  - Translation is simple phrase-based

- Purpose:
  
  - Have students get full experience
  
  - Find bugs/limitation in the system
  
  - Evaluation resulting systems for development time and accuracy