Speech Processing 15-492/18-492

Multilinguality
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
What you need

- **ASR**
  - Acoustic model (lots of speakers)
  - Pronunciation Lexicon
  - Language model

- **TTS**
  - Acoustic model (one speaker)
  - Pronunciation Lexicon
  - Text analysis
Writing Systems

- **Romanized writing systems**
  - *Latin-1 (iso-8599-1)*
  - Covers many Western Europeans languages
- **Cyrillic**
  - Covers many Eastern European Languages
- **Arabic Scripts**
  - Arabic(s), Farsi, Urdu, etc
- **Devenagari**
  - Covers many Northern India Languages
- **Chinese Hanzi**
  - Covers some Chinese dialects but different versions
- **Many other scripts some non-standard**
Writing Systems

- **Letter based**
  - *Latin, Cyrillic*

- **Consonant based**
  - *Arabic, Hebrew*

- **Mora based**
  - *Half syllable or syllable*
  - *Indian scripts, Japanese native scripts*

- **Syllable based**
  - *Hangul, Chinese*
Standards

- **Writing standards**
  - Taught at schools, newspapers, computer support
  - Typically standardized spelling
- **May be mostly spoken**
  - Occasionally written
Language Specific Issues

- **No explicit markings**
  - Stress, accent, tones

- **No word boundaries**
  - Chinese, Thai

- **No (short) vowels**
  - Arabic, Hebrew

- **Rich morphology**
  - Many different words in the languages
  - Finnish, Turkish, Greenlandic
Genre Specific Issues

- No capitals, punctuations
- Unpunctuated
- Plain vs polite form
- Speech vs text form
- Many foreign phrases
  - (technology directed genre’s)
- Many new abbreviations
  - E.g. SMS messages
Character Encoding

- **Unicode vs utf8 vs latin**
  - Documents mix them
- **Sometime accent omitted**
  - For ease of typing
- **Lots of standards**
  - Unicode, EUC, BIG5, TIS42, ...
  - Everyone has their own standard
- **Some create their own standards**
- **Mixed character sets**
Phoneme Sets

- Hard to find consensus for new languages
  - Typically lots of different dialects
- What level of distinction?
  - Some good for speech but not really phonetic
  - /t/ vs /dx/ in “water”
- Often doesn’t include foreign phones
  - /w/ in German is common for younger people
Words

- **May be hard to define**
  - No word boundaries

- **Rich morphology**
  - *Words have many variations of compounds*
  - *Yomenakatta -> could not read*
  - *Yomemasendeshita -> could not read (polite)*

- **Gender specific speech**
  - *Boku vs atashi*

- **Language mixtures**
Pronunciation lexicons

- “proper” speech vs “actual” speech
- Hard to generalize
  - Chinese
- Cross lingual pronunciations
  - “Human” (English/German)
“Industry” way

- Collect at least 300 hours of spoken speech
  - At least 20 different speakers
  - Mixture of gender, age, etc
  - Through desired channel (phone/desktop)
- Collect at least 5 hours from one speaker
  - High quality recording studio
- Data should be targeted to application
- Build pronunciation lexicon
  - Expert phonologist
Industry way

- Probably 3-6 months
  - Lead developer
  - Local language expert
  - Lots of human transcribers

- Costs?
  - Many hundreds of thousands
Or cheaper (?) ...

- Find existing data
  - Linguistic Data Consortium (UPenn)
  - ELRA (European equivalent)
  - Appen, Australia
  - Find local people who have collected data
- Found data might be in wrong format
  - Data cleaning is often the most expensive
Standardized Datasets

- **Global Phone**
  - 20+ languages, for ASR/TTS

- **LDC/DARPA/IARPA sets**
  - Mostly English, Arabic and Chinese

- **BABEL dataset**
  - 35 low resource languages (telephone conversations)

- **Librivox**
  - Audio books

- **Voxforge**
  - Open source collected languages

- **Mozilla**
  - Open source multilingual sets
CMU Wilderness Dataset

- 500+ Languages
  - 20 hours aligned for each language
  - Single speaker
  - Mined from read audio books (Bible)
  - 20+ languages, for ASR/TTS
Actual way

- Often mixture
  - Found data for initial model
  - Collect data with actual/initial application
Multilingual Systems

- **Support lots of different languages**
  - Press 1 for Spanish
  - Press 2 for Gujarati …

- **Automatically detect language**

- **Mixed language**
• **Speak in your language**
  - *Eki-mai no tsugi no bus no ha?*
  - *When is the next bus to the station*

• **Need multiple recognizers**
  - *Run in parallel and take best result*

• **Or shared acoustic models**
  - *Recognizing both languages at once (mix)*
- Code switching
  - European, India, Bilingual areas
  - Hinglish, Spanglish

- Borrowed words and phrases
  - Dad, time kyu hua hai
  - One lakh
  - Computer walla
  - numbers

- Can be inflected
  - Was updated -> up gedaten
Speech Processing 11-492/18-492

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

1. **Computerization**: Speech is key technology
   - Mobile Devices, Ubiquitous Information Access

2. **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 $L_s$

$AM_s$ $Dict_s$ $LM_s$ $Lex_{st}$ $LM_t$ $Dict_t$ $AM_t$

$L_s$ $L_{source}$ $L_{target}$

$L_t$
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 \( \rightarrow \) 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

input: Speech

AM

Lex

LM

NLP

/MT /

TTS

Output: Speech & Text

Hello
Input: Speech

Output: Speech & Text

---

Rapid Portability: Data

Phone set & Speech data

---

Hello

Input: Speech

AM
Lex
LM
NLP
MT
TTS

สวัสดี ครับ

Output: Speech & Text
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, formatting 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.

[Upload Audio]

Or, record audio: [Watch Demo Video]

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.co.cmu.edu:7000
GlobalPhone

**Multilingual Database**
- Widespread languages
- Native Speakers
- Uniform Data
- Broad Domain
- Large Text Resources
  - Internet, Newspaper

**Corpus**
- 19 Languages … counting
- \( \geq \) 1800 native speakers
- \( \geq \) 400 hrs Audio data
- Read Speech
- Filled pauses annotated

Now available from ELRA !!
Speech Recognition in 17 Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Word Error Rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese</td>
<td>10</td>
</tr>
<tr>
<td>German</td>
<td>11.8</td>
</tr>
<tr>
<td>English</td>
<td>14</td>
</tr>
<tr>
<td>Thai</td>
<td>14</td>
</tr>
<tr>
<td>Korean</td>
<td>14.5</td>
</tr>
<tr>
<td>Ch-Mandarin</td>
<td>14.5</td>
</tr>
<tr>
<td>Turkish</td>
<td>16.9</td>
</tr>
<tr>
<td>French</td>
<td>18</td>
</tr>
<tr>
<td>Portuguese</td>
<td>19</td>
</tr>
<tr>
<td>Croatian</td>
<td>20</td>
</tr>
<tr>
<td>Spanish</td>
<td>20</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>29</td>
</tr>
<tr>
<td>Russian</td>
<td>33.5</td>
</tr>
<tr>
<td>Afrikaans</td>
<td>20</td>
</tr>
<tr>
<td>Chinese</td>
<td>21.7</td>
</tr>
<tr>
<td>Arabic</td>
<td>23.4</td>
</tr>
<tr>
<td>Iraqi</td>
<td>29</td>
</tr>
</tbody>
</table>
Rapid Portability: Acoustic Models

Phone set & Speech data

AM
Lex
LM
NLP / MT
TTS

Input: Speech

Hello

Output: Speech & Text
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
**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.

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>postalveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place</td>
<td>P</td>
<td>T</td>
<td>D</td>
<td>ɹ</td>
<td>R</td>
<td>k</td>
<td>k</td>
<td>q</td>
<td>q’</td>
<td></td>
<td>ʔ</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>t</td>
<td>d</td>
<td>ɹ̚</td>
<td>r̚</td>
<td>ḵ</td>
<td>ḵ’</td>
<td>q̱</td>
<td>q’̱</td>
<td></td>
<td>ʔ̱</td>
</tr>
</tbody>
</table>

Nasal

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>N</th>
<th>N̩</th>
<th>N̟</th>
<th>n</th>
<th>n̩</th>
<th>n̟</th>
<th>n̩</th>
<th>n̟</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m</td>
<td>n</td>
<td>n̩</td>
<td>n̟</td>
<td>n</td>
<td>n̩</td>
<td>n̟</td>
<td>n̩</td>
<td>n̟</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Trill

<table>
<thead>
<tr>
<th></th>
<th>r</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r̄</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rapid Portability: Acoustic Model

Word Error rate [%]

Ø Tree  ML-Tree  Po-Tree  PDTS

0  0:15  0:15  0:25  0:25  0:25  1:30  16:30

69.1  57.1  49.9  40.6  32.8  28.9  19.6  19

Brazilian flag
Rapid Portability: Pronunciation Dictionary

**Pronunciation rules**

- "adios" → /a/ /d/ /i/ /o/ /s/
- "Hallo" → /h/ /a/ /l/ /o/
- 现在广播完了 → ???

**Input:** Speech

- AM
- Lex
- LM

**Output:** Speech & Text

- NLP
- MT
- TTS

**Languages supported:**

- العربية
- català
- 中国话
- hrvatski
- česky
- इंग्लिश
- ελληνικά
- עברית
- हिंदी
- italiano
- 日本語
- 한국어
- românește
- русский
- српски
- ภาษาไทย

**Flags:**

- China
- France
- India
- Germany
- Brazil
- Spain
- Russia

- United States
- Germany
- Brazil
- Spain
- Sweden
- Denmark

**Examples:**

- "adios" → /a/ /d/ /i/ /o/ /s/
- "Hallo" → /h/ /a/ /l/ /o/
- 现在广播完了 → ???

**Hello**
Problem:
• 1 Grapheme ≠ 1 Phoneme

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

Phoneme- vs Grapheme based ASR

The chart shows error rates for Phoneme, Grapheme, and Grapheme (FTT) for English, Spanish, German, Russian, and Thai languages. The error rates range from 0% to 50%. The Phoneme-based ASR shows lower error rates compared to the Grapheme-based ASR, with the Grapheme (FTT) showing further improvements.
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**  [Logout]

**Lexicon pronunciation creation**

**Rule entry**

3.0075187969925% Finished new word:

**at**

system suggested pronunciation: **AX T**  
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**
User: awb  Language: eng  Project: aug19  [Logout]

Lexicon pronunciation creation

Rule entry

3.5087719298246% Finished

new word: Jeanne

system suggested pronunciation: * AX N N  listen to it

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
dispersion (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:

Internet / TV

Inquiry

Automatic Extraction

LM

Bridge Languages

Text data

NLP / MT

TTS

Input: Speech

AM
Lex
LM

Output: Speech & Text

Hello

Hi /h//ai/
you /j/u/
we /w///i/

I am
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 distribution across tasks:**

- **Data:** AM (ASR) 5, Lex 3, LM (ASR, MT) 11
- **Training:** AM (ASR) 8, Lex 3, LM (ASR, MT) 11
- **Tuning:** AM (ASR) 7, LM (ASR, MT) 7
- **Evaluation:** TTS 5, S-2-S 5
- **Prototype:** TTS 5, S-2-S 5
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