Speech Processing for Unwritten Languages

Alan W Black
Language Technologies Institute
Carnegie Mellon University

ISCSLP 2016 – Tianjin, China
Speech Processing for Unwritten Languages

Joint work with
Alok Parlikar, Sukhada Parkar,
Sunayana Sitaram, Yun-Nung (Vivian) Chen,
Gopala Anumanchipalli, Andrew Wilkinson,
Tianchen Zhao, Prasanna Muthukumar.

Language Technologies Institute
Carnegie Mellon University

ISCSLP 2016 – Tianjin, China
Speech Processing

- The major technologies:
  - Speech-to-Text
  - Text-to-Speech
- Speech processing is text centric
Overview

- **Speech is not spoken text**
- **With no text what can we do?**
  - *Text-to-speech without the text*
  - *Speech-to-Speech translation without text*
  - *Dialog systems for unwritten languages*
- **Future speech processing models**
Most languages are not written

- Literacy is often in another language
  - e.g. Mandarin, Spanish, MSA, Hindi
  - vs, Shanghaiese, Quechua, Iraqi, Gujarati

Most writing systems aren’t very appropriate

- Latin for English
- Kanji for Japanese
- Arabic script for Persian
Writing Speech

- Writing is not for speech, it's for writing.
- Writing speech requires (over) normalization:
  - "gonna" → "going to"
  - "I'll" → "I will"
  - "John's late" → "John is late"
- Literacy is often in a different language:
  - Most speakers of Tamil, Telugu, Kannada write more in English than in their native language.
- Can try to force people to write speech:
  - Will be noisy, won't be standardized.
Less well-written language processing

Not so well defined
  - No existing resources (or ill-defined resources)
  - Spelling is not-well defined

Phoneme set
  - Might not be dialect appropriate (or archaic)
  - (Wikipedia isn't always comprehensive)

But what if you have (bad) writing and audio
  - Writing and Audio
Grapheme Based Synthesis

- Statistical Parametric Synthesis
  - More robust to error
  - Better sharing of data
  - Less instance errors
- From ARCTIC (one hour) databases (clustergen)
  - This is a pen
  - We went to the church and Christmas
  - Festival Introduction
# Other Languages

- Raw graphemes (G)
- Graphemes with phonetic features (G+PF)
- Full knowledge (Full)

<table>
<thead>
<tr>
<th>Language</th>
<th>G</th>
<th>G+PF</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>5.23</td>
<td>5.11</td>
<td>4.79</td>
</tr>
<tr>
<td>German</td>
<td>4.72</td>
<td>4.30</td>
<td>4.15</td>
</tr>
<tr>
<td>Inupiaq</td>
<td>4.79</td>
<td>4.70</td>
<td></td>
</tr>
<tr>
<td>Konkani</td>
<td>5.99</td>
<td>5.90</td>
<td></td>
</tr>
</tbody>
</table>

Mel-cepstral Distortion (MCD) lower is better
Unitran: Unicode phone mapping

- Unitran (Sproat)
  - Mapping for all unicode characters to phoneme
  - (well almost all, we added Latin++)
  - Big table (and some context rules)
    - Grapheme to SAMPA phone(s)
  - (Doesn't include CJK)
  - Does cover all other major alphabets
More Languages

- Raw graphemes
- Graphemes with phonetic features (Unitran)
- Full knowledge

<table>
<thead>
<tr>
<th>Language</th>
<th>G</th>
<th>Unitran</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi</td>
<td>5.10</td>
<td>5.05</td>
<td>4.94</td>
</tr>
<tr>
<td>Iraqi</td>
<td>4.77</td>
<td>4.72</td>
<td>4.62</td>
</tr>
<tr>
<td>Russian</td>
<td>5.13</td>
<td>4.78</td>
<td></td>
</tr>
<tr>
<td>Tamil</td>
<td>5.10</td>
<td>5.04</td>
<td>4.90</td>
</tr>
<tr>
<td>Dari</td>
<td>4.78</td>
<td>4.72</td>
<td></td>
</tr>
</tbody>
</table>
Let’s derive a writing system

• Use cross-lingual phonetic decoding
• Use appropriate phonetic language model

Evaluate the derived writing with TTS

• Build a synthesizer with the new writing
• Test synthesis of strings in that writing
Deriving Writing

Speech without Orthography → ASR: Phonetic Decoding → Phonetic Transcriptions → TTS Training

Phonetic Language Model (related language with orthography)
Acoustic Model (high resource language)

TTS Voice
Cross Lingual Phonetic Labeling

- For German audio
  - AM: English (WSJ)
  - LM: English
  - Example:

- For English audio
  - AM: Indic (IIIT)
  - LM: German
  - Example:
Iterative Decoding

Diagram:
- Speech without Orthography
  - ASR: Phonetic Decoding
    - Phonetic Transcriptions
      - TTS Training
      - TTS Voice
      - MCD Improved?
    - Acoustic Model Training (Yes)
    - Acoustic Model
  - Phonetic Language Model (related language with orthography)

Iterative Decoding: German

AM: English (WSJ)  
LM: English

Chart showing MCD values across iterations.
Find better Phonetic Units

- Segment with cross lingual phonetic ASR
- Label data with Articulatory Features
  - (IPA phonetic features)
- Re-cluster with AFs
• 26 streams of AFs
• Train Neural Networks to predict them
  • Will work on unlabeled data
• Train on WSJ (Large amount English data)
ASR: “Articulatory” Features

- These seem to discriminate better
Cluster New “Inferred Phones”
Synthesis with IPs

The diagram shows the MCD scores plotted against the number of inferred phones. The x-axis represents the number of inferred phones, ranging from 20 to 220, while the y-axis represents the MCD scores, ranging from 5.4 to 6.1. The line graph includes multiple lines, each representing different languages and their respective methods.

- **Hindi with IPs**
- **Hindi with ASR phones**
- **Dari with IPs**
- **Dari with ASR phones**
- **Iraqi with IPs**
- **Iraqi with ASR phones**

Each line represents the performance of synthesis with and without ASR phones for Hindi, Dari, and Iraqi languages.
IP are just symbols

• IPs don't mean anything
  • But we have AF data for each IP
  • Calculate mean AF value for each IP type
    • Voicing, Place of articulation ...
  • IP type plus mean/var AFs
Synthesis with IP and AFs

![Graph showing MCD scores vs. Number of Inferred Phones for different conditions: Hindi with IPs, Best previous, Dari with IPs, and Best previous.](image)
Need to find “words”

• From phone streams to words
  ● Phonetic variation
  ● No boundaries

• Basic search space
  ● Syllable definitions (lower bound)
  ● SPAM (Accent Groups) (upper bound)
  ● Deriving words (e.g. Goldwater et al)
Other phenomena

• But it's not just phonemes and intonation
  • Stress (and stress shifting)
  • Tones (and tone sondhi)
  • Syllable/Stress timing
  • Co-articulation
  • Others?

• [ phrasing, part of speech, and intonation ]
• MCD might not be sensitive enough for these
  • Other objective (and subjective measures)
• Method to derive new “writing” system
• It is sufficient to represent speech
• But who is going to write it?
Speech to Speech Translation

• From high resource language
  • To low resource language

• Conventional S2S systems
  • ASR -> text -> MT -> text -> TTS

• Proposed S2S system
  • ASR -> derived text -> MT -> text -> TTS
Audio Speech Translations

- From audio in target language to text in another:
  - Low resources language (audio only)
  - Transcription in high resource language (text only)
- For example
  - Audio in Shanghaiese, Translation/Transcription in Mandarin
  - Audio in Konkani, Translation/Transcription in Hindi
  - Audio in Iraqi Dialect, Translation/Transcription in MSA
- How to collect such data
  - Find bilingual speakers
  - Prompt in high resource language
  - Record in target language
Collecting Translation Data

- Translated language not same as native language
- Words (influenced by English) (Telugu)
  - “doctor” → “Vaidhyudu”
  - “parking validation” → “???”
  - “brother” → “Older/younger brother”
- Prompt semantics might changes
  - Answer to “Are you in our system?”
  - Unnanu/Lenu (for “yes”/”no”)
  - Answer to “Do you have a pen?”
  - Undi/Ledu (for “yes”/”no”)
Can’t easily collect enough data
  - Use existing parallel data and pretend one is unwritten
  - But most parallel data is text to text

Let’s pretend English is a poorly written language
Spanish -> English translation
  - But we need audio for English
  - 400K parallel text en-es (Europarl)

Generate English Audio
  - Not from speakers (they didn’t want to do it)
  - Synthesize English text with 8 different voices
  - Speech in English, Text in Spanish

Use “universal” phone recognizer on English Speech
  - Method 1: Actual Phones (derived from text)
  - Method 2: ASR phones
<table>
<thead>
<tr>
<th></th>
<th>I declare resumed the session of the European Parliament adjourned on . . .</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>ay d ih k l eh r r ih z uw m d dh ax s eh sh ax n aa v dh ax y uh r ax p iy ax n p aa r l ax m ax n t ax jh er n d aa n . . .</td>
</tr>
</tbody>
</table>
Phone to “words”

- Raw phones too different to Target (translation) words
  - Reordering may happen at phone level
- Can we cluster phone sequences as “words”
  - Syllable based
  - Frequent n-grams
  - Jointly optimize local and global subsequences
    - Sharon Goldwater (Princeton/Edinburgh)
- “words” do not need to be source language words
  - “of the” can be a word too (it is in other languages)
### Table 2: Examples with naïve clustering

<table>
<thead>
<tr>
<th>Method 1</th>
<th>ay</th>
<th>d_ih</th>
<th>k_l_eh</th>
<th>r_r_ih</th>
<th>z_uw</th>
<th>m_d_dh</th>
<th>a_x_s_eh</th>
<th>sh_a_x_n</th>
<th>a_a</th>
<th>v_dh_a_x_y</th>
<th>u_h</th>
<th>r_a_x_p</th>
<th>i_y</th>
<th>a_x</th>
<th>n_p_a_a_r_l</th>
<th>a_x</th>
<th>m_a_x</th>
<th>n_t_a_x_jh</th>
<th>e_r</th>
<th>n_d_a_a_n_f_r</th>
<th>. . .</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 2</td>
<td>AY</td>
<td>D_IH</td>
<td>K_L_EH</td>
<td>R_IY</td>
<td>Z.D_UW</td>
<td>I_H</td>
<td>N_G_DIH_S</td>
<td>A_E</td>
<td>S_H_A_H_N</td>
<td>A_H</td>
<td>V_DH_A_E_T_Y</td>
<td>A_O</td>
<td>R_P_IY</td>
<td>A_E</td>
<td>N_D_P_A_A_R_T_L</td>
<td>I_H</td>
<td>M_A_E</td>
<td>N_D_I_H_JH</td>
<td>E_R</td>
<td>N_D_A_A_N_F_R</td>
<td>. . .</td>
</tr>
</tbody>
</table>

*English: phones to syls*
### Table 3: Examples with most-frequent-ngrams clustering

| Method 1 | ay_d   ih_k_l_eh_r   r   ih_z   uw_m   d   dh_ax_s_eh_sh_ax_n   aa_v_dh_ax   y_uh_r_ax_p_iy_ax_n   p_aa_r_l_ax_m_ax_n_t_ax_jh   er   n_d   aa_n ... |
|----------|---------|-----------------------------|------------------|------------------|----------------|---------------|-----------------------------|----------------|-------------------|------------------|------------------|------------------|---------------|-----------------------------|---------------|
| Method 2 | AY_D   IH_K_L   EH_R_IY_Z   D   UW_IH_NG   DH_IH_S   AE_SH   AH_N   AH_V_DH_AE_T   Y_AO_R   P_IY   AE_N_D   P_AA_R_T   L   IH_M_AE_N_D   IH_JH   ER   N_D   AA_N ... |
Table 4: Examples with Goldwater clustering

<table>
<thead>
<tr>
<th>Method 1</th>
<th>aydihklehr rihzuwmddhaxseh shaxnaav dhaxyuhraxpiyaxn paarlaxmaxnt axjh-ernd aanfraydiy …</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 2</td>
<td>AYDIHKL EHRIYZ DUWI-HNGDHIHS AESHAHNAHV DHAETYAORPIY AEND PAARTLIHM AEND IHJHERN D AAN …</td>
</tr>
<tr>
<td>Method</td>
<td>Words</td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
</tr>
<tr>
<td>Oracle</td>
<td>35.76</td>
</tr>
<tr>
<td>Method 1</td>
<td>20.45</td>
</tr>
<tr>
<td>Method 2</td>
<td>13.81</td>
</tr>
</tbody>
</table>
300K parallel sentences (FBIS)
- Chinese synthesized with one voice
- Recognized with ASR phone decoder

<table>
<thead>
<tr>
<th>English gloss</th>
<th>an international audience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word (hanzi)</td>
<td>国际 视听</td>
</tr>
<tr>
<td>Word (pinyin)</td>
<td>guójì shìtīng</td>
</tr>
<tr>
<td>Syllable (pinyin)</td>
<td>guó jì shì tīng</td>
</tr>
<tr>
<td>Phone (pinyin)</td>
<td>guó jì shì tīng</td>
</tr>
<tr>
<td>Goldwater (pinyin)</td>
<td>guójì shītīng</td>
</tr>
<tr>
<td>Phone (ASR)</td>
<td>K L IH K S IY SH IY EY T S L IH M P</td>
</tr>
<tr>
<td>Goldwater (ASR)</td>
<td>KLIHKSIY SHIYEYT S LIHMP</td>
</tr>
<tr>
<td></td>
<td>Word</td>
</tr>
<tr>
<td>---------------</td>
<td>-------</td>
</tr>
<tr>
<td>Hanzi</td>
<td>29.05</td>
</tr>
<tr>
<td>Pinyin</td>
<td>28.98</td>
</tr>
<tr>
<td>Pinyin (toneless)</td>
<td>28.30</td>
</tr>
<tr>
<td>ASR</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Can we interpret unwritten languages
- Audio -> phones -> “words”
- Symbolic representation of speech

SDS for unwritten languages:
- SDS through translation
  - Konkani to Hindi S2S: + conventional SDS
- SDS as end-to-end interpretation
  - Konkani to symbolic: + classifier for interpretation
But speech is speech not text

What about conversational speech

Laughs, back channels, hesitations etc

Do not have good textual representation

Larger chunks allow translation/interpretation
Phonet for Unwritten Languages

- Phonetic representation from acoustics
  - Cross lingual, phonetic discovery
- Word representation from phonetic string
  - Larger chunks allow translation/interpretation
- Higher level linguistic function
  - Word classes (embeddings)
  - Phrasing
  - Intonation
Conclusions

- Unwritten languages are common
- They require interpretation
- Can create useful symbol representations
  - Phonetics, words, intonation, interpretation
- Let’s start processing speech as speech