TTS without the Text: Speech Synthesis in Underresource Languages

Alan W Black and Alok Parlikar, Sukhada Palkar, Sunayana Sitaram, Yun Nung Chen, Gopala Anumanchipalli

Language Technologies Institute
Carnegie Mellon University
Building Synthetic Voices

- **Festvox Toolkit**
  - *All the tools, all the meta-tools*

- **CMU Spice**
  - *Web-based interface (for ASR and TTS)*

- **Text and Audio only**
  - *With limited knowledge of language*

- **Audio only**
  - *(Most languages) ...*
Core requirements

- Speech corpus
  - Single speaker, well recorded,
  - phonetically balanced

- Text analysis
  - Numbers, symbol etc mapping to words

- Pronunciation Lexicon
  - Mapping of words to pronunciation
  - LTS for unknown words

- Phoneme Set definition
  - Other factors, stress, tone accent

- High-level NLP
  - Part of Speech Tagging, “Parsing”
Language Types

- English: Few Languages
- Marathi: Many Languages
- Konkani: Most Languages
Special Case: Audio Only

- Audio but with unaligned text
  - In high resource language
- Splitting into utterance sized chunks
  - Interslice tool (but requires text)
  - Good for speeches, audio books etc
- Audio only (no transcription)
  - Hand transcribe it (not very reliable)
  - Use (inappropriate) ASR
  - Original
  - ASR synth
  - Full synth
In a study of 72 ballot issues in Massachusetts and three other states, Boston University Political Science Professor, Betty Zisk, found that 88% of the battles were won by the side that spent the most money.

In Estonian, 72 validations in Massachusetts in three other states Boston University Political Science Professor Buddy says about 80% of the battles won by the side could spend the most money.

Synthesis of actual text with “ASR” transcripts
Low Resource languages

- Not so well defined
  - No existing resources (or ill-defined resources)
- Phoneme set
  - Might not be dialect appropriate (or archaic)
  - (Wikipedia isn't always comprehensive)
- Only have
  - Writing and Audio
Naive Grapheme Based Synthesis

◆ “phonemes” are “letters”
  ● alan → a l a n
  ● black → b l a c k
  ● But wont work for English ... (?)

◆ From ARCTIC (one hour) databases (unit selection)
  ● This is a pen
  ● We went to the church and Christmas
  ● Festival Introduction
  ● do eight meat
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**
- **Graphemes with phonetic features**
- **Full knowledge**

<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)
  - Big table (and some context rules)
  - Doesn't include CJK or Latin Alphabet
    - We added support for Latin Alphabet
  - 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>
<tr>
<td>Thai</td>
<td>4.82</td>
<td>4.98</td>
<td></td>
</tr>
</tbody>
</table>
Beyond Phonemes

• **Speech is not just strings of phonemes**
• **Prosodic Modeling**
  - *Duration modeling*
  - *Intonation*
  - *Phrase Breaks*
“Grammar” Base Approach

• “There are five hundred people in the room”
• “There are BB five hundred people BB in the room”
• ((There are) (five hundred people) (in the room))
• ((EX VBP) (CD CD NNS) (IN DT NN)))

• Train Stochastic Context Free Grammar on this
  • “acoustic syntax”
• Define “syntactic” feature for CART

[Parlikar PhD 2013]
LR Phrase Break Model
Intonation

- "accents" (deviations from default declination)
- F0 contour itself

Statistical Phrase Accent Modeling

- Find F0 contour
- Find accent groups
  - That can be best parameterized
- Multi-level decomposition
  - Phrase, Accent, Micro-prosody)
Multi-level F0 Modeling

- Original F0
- Extracted Phrase Component
- Extracted Accent Component

log (F0) vs Time
Accent Groups

• From words
  ● Find best accent grouping for Spam model
  ● Build “grammar” and induce SCFG
• Much preferred and more “natural”
  ● Requires no explicit accent type design

[Anumanchipalli PhD 2013]
Most languages have no (good) orthography
- Inappropriate alphabet
- No standardized spelling
- Not the language literacy is taught in

Try writing your dialect down

Language without text
- More likely to require speech technology
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: German

AM: English (WSJ)  
LM: English

Graph showing iteration vs. MCD.
Find better Phonetic Units

• Use cross lingual phonetic ASR to get segments
• Label data with Articulatory Features
  • (IPA phonetic features)
• Find new set of units based on Afs
Articulatory Features (Metze)

• 26 streams of Afs
• Train Neural Networks to predict them
  • Will work on unlabeled data
• Train on WSJ (Large amount English data)
These seem to discriminate better
Cluster New “Inferred Phones”
Synthesis with IP
IPs don't mean anything
  • But we have AF data for each IP
  • So calculate mean AF value for each IP type
    • Voicing, Place of articulations ...
Synthesis with IP and AFs
German (Oracle)
Need to find “words”

• From phone streams to words
  ● Phonetic variation
  ● No boundaries
• Basic direction
  ● 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?

• MCD might not be sensitive enough for these
  • Other objective (and subjective measures)
What do you synthesize

• But what is the input ....

• Usage scenarios
  • Require bus Information in Konkani
  • Have working TTS in Hindi
    • Record Konkani with Hindi prompts
    • Build (new) TTS engine for Konkani
    • Use SMT to translation Hindi to “Konkani”

• TTS engine
  Input: Hindi text
  Output: Konkani speech
To do ...

• Build cross lingual TTS engine
  - Hindi in, Konkani out
  - Issues: recording, SMT etc
• How much data
  - About an hour of speech (?)
  - But current test sets are (too) well recorded
• Better cross lingual labeling
  - Articulatory Feature labeling (Metze)
• ASR part ...