Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language

Presenter:
Dr. Tania Habib
Outline:

• Overview
• Unit selection vs HMM based Speech Synthesis (HTS) [1]
• Development
• Requirements for Voice building
• Data Set
• Challenges
• Subjective Evaluation
• Erroneous Words
• Summary
Speech Synthesis Overview:

Text to be Synthesized

Natural Language Processing (NLP)

Speech Synthesis Engine

Synthesized Speech

Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language

CLT14 | Center for Language Engineering (CLE), 2014
Types of Speech Synthesis:

• Rule-based, *formant synthesis*
  • Hand-crafting each phonetic units by rules

• Corpus based:
  • *Concatenative synthesis* (Unit Selection)
    • High quality speech can be synthesized using waveform concatenation algorithms.
    • To obtain various voices, a large amount of speech data is necessary.

• *Statistical parametric synthesis* (HMM based)
  • Generate speech parameters from statistical models
  • Voice quality can easily be changed by transforming HMM parameters.
## Unit Selection vs. HTS

<table>
<thead>
<tr>
<th></th>
<th>Unit Selection</th>
<th>HTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advantages:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Quality</td>
<td>High Quality</td>
<td></td>
</tr>
<tr>
<td>at Waveform level</td>
<td>at Waveform</td>
<td>Small Foot Print</td>
</tr>
<tr>
<td>(Specific Domain)</td>
<td>level (Specific Domain)</td>
<td>Smooth</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stable Quality</td>
</tr>
<tr>
<td><strong>Disadvantages:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Large footprints</td>
<td>• Large</td>
<td>Vocoder sound</td>
</tr>
<tr>
<td>• Discontinuous</td>
<td>footprints</td>
<td>(Domain-independent)</td>
</tr>
<tr>
<td>• Unstable quality</td>
<td>• Discontinuous</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HTS Overview:

Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language

SPEECH DATABASE

Speech signal

Excitation Parameter extraction

Spectral Parameter Extraction

Excitation parameters

Spectral parameters

Training HMMs

Parameter generation from HMMs

Context-dependent HMMs & state duration models

Excitation parameters

Spectral parameters

Excitation generation

Synthesis filter

SYNTHESIZED SPEECH

Text analysis

Labels

Labels

TEXT

Training part

Synthesis part

[HTS Slides released by HTS Working Group slide no. 21]
Preliminary requirements for the HTS toolkit:

1. Annotated Training data.
2. Define speech features (MFCC, F0 and duration) for model training.
3. Sorting out unique context-dependent as well as context-independent phonemes (from the training data) for model training.
4. Unified question file for spectral, F0 and duration for context clustering.
Data Set Used:

• Source:
  • Paragraphs taken from Urdu Qaida of Grade 2 and 4 respectively

• Duration:
  • 30 minutes

• Total number of utterances:
  • 347

• Recording parameters:
  • Sample rate: 8KHz (up-sampled to 48KHz)
  • Channel: Mono
  • Recording format: .WAV
  • Speaker: Native Urdu female speaker
Challenges:

• Generation of the full-context style labels.
• Addition of Prosodic Layers
  • Segment
  • Stress
  • Syllable
  • Word
• Unbalanced Training Data
• Defining the Question Set (Context Clustering)
Full-Context Format (1/2):

Segmental Context:

- Current Phoneme
- Previous two Phonemes
- Next two Phonemes

Supra-Segmental Context:

- Syllable
- Stress
- Word
- Phrase
- POS

Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language
Full-Context Format (2/2):

x^x-SIL+A=L@1_0/A:0_0_0/B:0-0-0@1-0&1-1#1-1$1-1!0-0;0- ...
x^SIL-A+L=I_l@1_1/A:0_0_0/B:0-0-1@1-2&1-9#1-3$1-1!0-2;0- ...
SIL^A-L+I_l=A@1_2/A:0_0_1/B:0-0-2@2-1&2-8#1-3$1-1!0-1;0-0 ...
A^L-I_l+A=P@2_1/A:0_0_1/B:0-0-2@2-1&2-8#1-3$1-1!0-1;0- ...

Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language
Questions on Segmental/Prosodic Layers:

Phoneme
- \{preceding, succeeding\} two phonemes
- current phoneme

Syllable
- \#\ of phonemes at \{preceding, current, succeeding\} syllable
- \{accent, stress\} of \{preceding, current, succeeding\} syllable
- Position of current syllable in current word
- \#\ of \{preceding, succeeding\} \{accented, stressed\} syllable in current phrase
- \#\ of syllables \{from previous, to next\} \{accented, stressed\} syllable
- Vowel within current syllable

Word
- Part of speech of \{preceding, current, succeeding\} word
- \#\ of syllables in \{preceding, current, succeeding\} word
- Position of current word in current phrase
- \#\ of \{preceding, succeeding\} content words in current phrase
- \#\ of words \{from previous, to next\} content word

Phrase
- \#\ of syllables in \{preceding, current, succeeding\} phrase

Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language
Addition of Stress/Syllable Layer:

- Added layers:
  - Stress
  - Syllable

[Image of graph showing phonetic transcription and waveform]
Unbalanced Training data:

- High occurrence for vowels
- Some of the phonemes were completely ignored

Figure. Phoneme Coverage for the 30-min speech data
Context Clustering (Question Set):

• Number of possible combinations are quite enormous with 53 different questions.

• Possible contexts = \( C^n \)
  
  where \( C \) = Total count of basic phonetic units,
  
  \( n \) = Total number of Questions

  • With only Segmental Context (n=5) Possible models are:

  \[ 66^5 \approx 1252 \text{ million} \]

• If we consider all the context, it will be practically infinite.

Solution:

• Record data having maximum phoneme coverage at tri-phone or di-phone level.

• Apply context clustering technique to classify and share acoustically similar models
Subjective Evaluation:

• Testing Methodology:
  • Mean Opinion Score (MOS)[3] for:
    • Naturalness
    • Intelligibility

• Naturalness:

  How close it seems to be produced by a human?

• Intelligibility:

  How much conveniently the word was recognized?
Subjective Testing (Results):

<table>
<thead>
<tr>
<th>Listener Type</th>
<th>MOS Naturalness</th>
<th>MOS Intelligibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical 1</td>
<td>3.23</td>
<td>3.65</td>
</tr>
<tr>
<td>Linguistic 1</td>
<td>2.82</td>
<td>3.66</td>
</tr>
<tr>
<td>Linguistic 2</td>
<td>2.86</td>
<td>3.58</td>
</tr>
<tr>
<td>Linguistic 3</td>
<td>3.48</td>
<td>3.52</td>
</tr>
</tbody>
</table>

Table 1. Mean Opinion Score (MOS) results of four listeners
Erroneous words:

<table>
<thead>
<tr>
<th>Nastalique Style</th>
<th>CISAMPA (Correct)</th>
<th>Listened (Incorrect)</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>طرف</td>
<td>T_DARAF</td>
<td>T_DALAF</td>
<td>5.92</td>
</tr>
<tr>
<td>گا</td>
<td>GA_A</td>
<td>D_DA_A</td>
<td>1.35</td>
</tr>
<tr>
<td>معلوم</td>
<td>MAYLU_UM</td>
<td>MAT_DLU_UM</td>
<td>0.00</td>
</tr>
<tr>
<td>تهی</td>
<td>T_D_HA_Y</td>
<td>T_SA_Y</td>
<td>0.66</td>
</tr>
<tr>
<td>رزی</td>
<td>RAZI_I</td>
<td>RAD_DI_I</td>
<td>0.88</td>
</tr>
<tr>
<td>کیونک</td>
<td>KIU_U_NKA_Y</td>
<td>T_SU_NKA_Y</td>
<td>0.15</td>
</tr>
<tr>
<td>حق</td>
<td>HAQ</td>
<td>HABS</td>
<td>0.46</td>
</tr>
<tr>
<td>بعد</td>
<td>BAYD_D</td>
<td>BAD_D</td>
<td>0.00</td>
</tr>
<tr>
<td>خیال</td>
<td>XAJA_AL</td>
<td>FIJA_AL</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Table 2. Synthesized words with errors
Some Synthesized Examples:

Synthesized:                  Training Set:

Seen Context: 

Un-seen Context: 

Different Carrier Word: 

Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language

CLT14 Center for Language Engineering (CLE), 2014
Summary:

- **Text to Speech Synthesis (TTS):**
  - Concatenative
  - Parametric (Hmm based)

- **Requirement for Voice building**
  - Annotated speech corpus
  - Speech features
  - Question file

- **Challenges**
  - Full context style labels
  - Addition of prosodic layers
  - Question file for context clustering

- **Subjective Evaluation**

- **Erroneous words**
References:

