
Presenter:

Omer Nawaz
Research Officer (III)
Speech Synthesis Overview:

- Text to be Synthesized
- Natural Language Processing (NLP)
- Speech Synthesis Engine
- Synthesized Speech
Introduction:

Rule-based, *formant synthesis*

- Hand-crafting each phonetic units by rules

CORPUS-BASED:

• *Concatenative synthesis*
  - 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*
  - Generate speech parameters from statistical models
  - Voice quality can easily be changed by transforming HMM parameters.
Approaches at CLE:

CORPUS-BASED:
• *Unit Selection*
• *HMM based.*

Comparison of two Approaches:

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

Source Filter Model:

- **Source excitation part**
  - Pulse train
  - White noise

- **Vocal tract resonance part**
  - Excitation $e(n)$
  - Linear time-invariant system $h(n)$

- Speech $x(n) = h(n) \ast e(n)$

- The $h(n)$ is defined by the state output vector of the HMM e.g. mel-cepstrum
General Overview (HTS):

Speech Input

Labels

Extract Spectrum, $F_0$, labels

Train Acoustic Models

Parameter Generation

Text Input

Synthesis Filter

Synthesized Speech

Stored Models

Training Part

Synthesis Part
Challenges:

Generation of the full-context style labels.
Addition of Stress/Syllable Layer.
Defining the Question Set.
I-Context Label Style:

- **Phoneme sequence**

  - Tri-phone context dependent model

  - Full-context style context dependent model
I-Context Format:

\[ x^{SIL} + A = L \]

\[ SIL^{A} + L = I \]

\[ L^{A} + I = A \]

\[ L^{A} + I = A \]

على الرغم من أنني لا أستطيع قراءة النص العربي بشكل طبيعي، إلا أنني أستطيع قراءة النص الإنجليزي بشكل طبيعي.
I-Context Format:

<table>
<thead>
<tr>
<th>Segmental</th>
<th>Supra-Segmental</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Current Phoneme</td>
<td>- Syllable</td>
</tr>
<tr>
<td>- Previous two Phonemes</td>
<td>- Stress</td>
</tr>
<tr>
<td>- Next two Phonemes</td>
<td>- Word</td>
</tr>
<tr>
<td></td>
<td>- Phrase</td>
</tr>
<tr>
<td></td>
<td>- POS</td>
</tr>
</tbody>
</table>
Steps to Generate Full-Context Labels:

1. Extract Segmental & Word Layer
2. Apply Stress & Syllabification Rules
3. Align Syllable Boundaries with Segmental Layer
4. Generate new TextGrid File with Additional Layers
5. Convert to Full-Context format
TextGrid Format:
Steps to Generate Full-Context Labels:

1. Extract Segmental & Word Layer
2. Apply Stress & Syllabification Rules
3. Align Syllable Boundaries with Segmental Layer
4. Generate new TextGrid File with Additional Layers
5. Convert to Full-Context format
tGrid Format with Additional Layers:
Context Clustering (Question Set) 1/2:

Number of possible combinations are quite enormous with these 53 different contexts.

With only Segmental Context 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
Context Clustering (Question Set) 2/2:

Phoneme

- {preceding, current, succeeding} phonemes

Stress/Syllable/Word/

- # of phonemes at {preceding, current, succeeding} syllable
- stress of {preceding, current, succeeding} syllable
- Position of current syllable in current word
- # of syllables {from previous, to next} stressed syllable
- Vowel within current syllable
- # of syllables in {preceding, current, succeeding} word
Some Synthesized Examples:

Seen Context:

Un-seen Context:

Different Carrier Word:

Training Set:
Questions ?