Pioneering Dzongkha Text To Speech Synthesis

Department Of Information and Technology, Bhutan. NECTEC, Thailand.
Overview

- Introduction
- Development
  - Phoneme Design for Dzongkha TTS
  - TTS Design and development
- Evaluation and discussion
- Future prospects
- Conclusion
It consisted of
- designing a phoneme set
- building a text processor
- designing and collecting speech database
- training HMM under HMM-based speech synthesis system (HTS) toolkit
- integrating all components in an application.
The key features of the TTS
- Text analysis
- Speech synthesis
- Figure 1 below shows these two features
- Text analysis finds intermediate forms (Syllables in case of Dzongkha TTS)
- Synthesizing generates speech signals from that intermediate form.
Figure 1: The common form model of TTS (P. Taylor. 2008)
Dzongkha TTS
  - HMM-based
    • Uses acoustics parameters to generate speech
    • These are synthesized from context dependent HMM-models
  - HTS version 2.0
  - MCEP (mel-cepstral coefficients)
  - Log F0
  - Duration parameters
Dzongkha sound system

Figure 4(a): Dzongkha IPA table (consonants)

<table>
<thead>
<tr>
<th>Place Manner</th>
<th>Labio</th>
<th>Labio-dental</th>
<th>Dental/alveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Laryngeal/G</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stops</strong></td>
<td>Voiceless</td>
<td>P(p) (ʰ)</td>
<td>T(t) (ʰ)</td>
<td>Tr (ʰ)</td>
<td>K(k) (ʰ)</td>
<td>A(ʔ) (ʰ)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aspirated</td>
<td>Ph(pʰ) (ʰ)</td>
<td>Th(tʰ) (ʰ)</td>
<td>Thr (ʰ)</td>
<td>Kh(kʰ) (ʰ)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Voiced</td>
<td>B(b) (ʰ)</td>
<td>D(d) (ʰ)</td>
<td>Dr (ʰ)</td>
<td>G(g) (ʰ)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fricatives</strong></td>
<td>Voiceless</td>
<td>Sa(s) (ʰ)</td>
<td>Sh(ʃ) (ʰ)</td>
<td>Ha(hh) (ʰ)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Voiced</td>
<td>z(z) (ʰ)</td>
<td>Zh(ʒ) (ʰ)</td>
<td>'A(ɦ) (ʰ)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Affricatives</strong></td>
<td>Voiceless</td>
<td>Ts(ts) (ʰ)</td>
<td>C(kʃ) (ʰ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aspirated</td>
<td>Tsh(tsʰ) (ʰ)</td>
<td>Ch(kʃʰ) (ʰ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Voiced</td>
<td>Dz(dz) (ʰ)</td>
<td>J(dʒ) (ʰ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Trill</strong></td>
<td></td>
<td>R(r) (ʰ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lateral</strong></td>
<td></td>
<td>L(l) (ʰ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Approximant</strong></td>
<td>W(w) (ʰ)</td>
<td>Y(j) (ʰ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nasals</strong></td>
<td>M(m) (ʰ)</td>
<td>N(n) (ʰ)</td>
<td>Ny(ɲ) (ʰ)</td>
<td>Ng(ŋ) (ʰ)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dzongkha sound system

- Figure 4(b): IPA table for Dzongkha (vowels)
Dzongkha sound system

- Representation of spoken Dzongkha
  - Initial consonants
  - Consonant clusters with single consonants
  - Vowels
  - Diphthongs
  - An inherent vowel 'a' is always present with single consonants
  - Some vowels are modified when root letter combines with certain suffixes
  - Clusters in Dzongkha is represented by stacking root letter over the subjoined letter
Dzongkha sound system

- Dzongkha tone system
  - Two tone system
  - The low tone normally used
  - The high which is the modification of the low tone
  - Modification depends on combination of certain prefixes (ʼgil, ʼgil, ʼgil, ʼgil), head letter (ʼn, ʼň) and subjoined (ʼn) with root letter.
Dzongkha sound system

- Figure 5: Normalized F0 contour of the syllable 'lam' showing high tone (meaning monk) and low tone (meaning road or way)
Observations during transcription

- 30 initial consonants, 5 initial consonant clusters, 10 vowels and 10 diphthongs defined
- Single phonemes from figure 4 were employed
- Four more vowels were observed and defined separately (‘aa’, ‘ii’, ‘uu’, ‘oo’).
- Consonant clusters mostly formed by combination with 'r' sound
- Some suffixes are not pronounced (‘d’ 's' 'hh')
- Certain suffixes modifies the vowel
### Phoneme design for TTS

#### Table 1: Dzongkha phoneme inventory for TTS

<table>
<thead>
<tr>
<th>Type</th>
<th>Symbol (IPA/Computerized)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial consonant (Ci)</strong></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>k, kʰ/kh, g/g, ɲ/ŋ, tʃ/c, ʃʰ/ch, ɗʒ/j, ɲ/ny, t, tʰ/th, d, n, p, pʰ/ph, b, m, ts, tsʰ/tʃh, dz, w, ʒ/zh, z, hʰ/hh, j/y, i/r, l, ʃ/sh, s, h, ?/@</td>
</tr>
<tr>
<td>Cluster</td>
<td>d.r/dr, tɻ, tʰ/ɻh, lʰ/lh, ḥɻ/ɻl</td>
</tr>
<tr>
<td><strong>Vowel (V)</strong></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>a, i, u, e, o, ue, a:/aa, i:/ii, u:/uu, o:/oo</td>
</tr>
<tr>
<td>Diphthong</td>
<td>ai, au, ae, ui, oi, ou, eu, ei, eo, iu</td>
</tr>
<tr>
<td><strong>Final consonant (Cf)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>g/g, ɲ/ŋg, n, b, m, i/r, l, p</td>
</tr>
<tr>
<td><strong>Tone (T)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ɻ/0, ɻ/1</td>
</tr>
</tbody>
</table>
Phoneme design for TTS

- Vowel modification with suffices.
- Table 2: Modification of vowel.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>g</th>
<th>ng</th>
<th>n</th>
<th>b</th>
<th>m</th>
<th>r</th>
<th>l</th>
<th>p</th>
<th>d</th>
<th>s</th>
<th>hh</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>e</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>e</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>o</td>
<td>e</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>e</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Phoneme design for TTS

- Tonal representation
  - Digit symbol '0' for low
  - Digit symbol '1' for high
TTS design and development

- The system consists of two main modules, text analysis and speech synthesis.
- Figure 6: The proposed system structure.
TTS design and development

- **Text analysis**
  - Implemented using a dictionary based G2P
  - Presence of a syllable marker makes it easier to implement G2P using a look up dictionary

- **Dictionary**
  - A Dzongkha text corpus of 40,000 sentences were collected
  - Top 4000 distinct syllables occurring were included in the dictionary
TTS design and development

- **Speech synthesis**
  - A corpus of 509 sentence included
  - These had to cover all 53 phonemes including two tones shown above in Table 1.

- **The sentence selection**
  - Iteratively select a sentence with most distinctive tonal di-phones
  - Stop when all tonal di-phones in text corpus are included
TTS design and development

- Table 3: Dzongkha speech corpus statistics.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of sentences</td>
<td>509</td>
</tr>
<tr>
<td>No. of syllables</td>
<td>5,404</td>
</tr>
<tr>
<td>No. of tonal diphones</td>
<td>6,048</td>
</tr>
<tr>
<td>No. of distinct tonal diphones</td>
<td>539</td>
</tr>
</tbody>
</table>
TTS design and development

- Building synthesizer
  - Mel-Cestrum (MCEP), duration and Log fundamental frequency (Log F0) were extracted from each utterance in the speech corpus
  - By using HTS with HTK and SPTK HMMs can trained in a flat start manner
  - It doesn't require any phoneme boundary tag but only phoneme transcription of each utterances
TTS design and development

- A clustering tree designed for Dzongkha phoneme is used in HMM state tying.

  Figure 7: A part of clustering tree used for HMM state tying.
TTS design and development

- Building the synthesizer
  - Using the training script in the HTS, the HMMs are trained to construct the synthesizer
  - Given trained HMMs, the “hts-engine” command with the toolkit could be evoked to synthesize speech.
TTS design and development

- Figure 8: HTS toolkit usage.
Evaluation and discussion

- Evaluation
  - based on mean opinion score
  - Fifteen Bhutanese were asked to evaluate

- Score system
  - 1 to 5
  - 1 for worst
  - 5 for best

- Result
  - human speech rated 3.93
  - synthesized rated 3.19
Future prospects

- Enlarging speech corpus with larger di-phone coverage
- More distinct syllables required by the G2P module
- Important prosody generation modules
  - pausing between words and phrases
  - duration and F0 modeling
Conclusion

- Building the first Dzongkha TTS
  - designing a phoneme inventory
  - building a text processor
  - designing and creating the speech database
  - training HMMs under HTS frame work
  - integrating all these into an application
- Yet more work needs to be done to improve speech quality as mentioned in future prospects