Overview of NITECH HMM-based text-to-speech system for Blizzard Challenge 2014

Kei Sawada, Shinji Takaki, Kei Hashimoto, Keiichiro Oura, and Keiichi Tokuda
Nagoya Institute of Technology (NITECH)

Blizzard Challenge 2014 Workshop on Sep 19, 2014
Outline

Background
Blizzard Challenge 2014 rules
System overview
  • Speech recognizer (SR)
  • Word aligner (WA)
  • Speech synthesizer (SS)
  • Grapheme-to-phoneme (G2P) converter
Experiments
Conclusions
Background

Text-to-speech (TTS) system

- TTS have been used widely in various applications
  - Car navigation, mobile phone, spoken dialogue, etc.
- Main components of TTS system
  - Text analysis: lexicon

Blizzard Challenge [Black, et al.]

- Blizzard Challenge was started in order to better understand and compare research techniques

NITECH has participated using HMM-base TTS
Outline

Background
Blizzard Challenge 2014 rules
System overview
  • Speech recognizer (SR)
  • Word aligner (WA)
  • Speech synthesizer (SS)
  • Grapheme-to-phoneme (G2P) converter
Experiments
Conclusions
Blizzard Challenge 2014 rules

TTS systems of six Indian languages
• Assamese, Gujarati, Hindi, Rajasthani, Tamil, Telugu

Hub task (IH1)
• Build one voice TTS system in each Indian language
• Provided speech data and corresponding text

Spoke task (IH2)
• Build a multilingual TTS system (Indian and English)
• Training data for this task was same as for Hub task
• Sample input text (Hindi and English):

उन्हें 10 दिन तक rehab करना होगा और उसके बाद उनका fitness test लया जाएगा
Difficulty in TTS system building

Phonemeset of target Indian language doesn’t exist
- Use a speech recognizer of English
  - Obtain label sequences of target Indian language
  - Also useful for multilingual speech synthesis

Label sequence doesn’t include word breaking info.
- Use multigram word aligner
  - Obtain word breaking information of label sequence

Lexicon of target Indian language doesn’t exist
- Use joint multigram grapheme-to-phoneme converter
  - Obtain label sequences of given input text
Outline

Background
Blizzard Challenge 2014 rules

System overview

- Speech recognizer (SR)
- Word aligner (WA)
- Speech synthesizer (SS)
- Grapheme-to-phoneme (G2P) converter

Experiments

Conclusions
System overview

- **English database**
  - Training SR
  - Speech recognizer (initial)
  - Label (sentence)

- **Indian speech database**
  - Training SR
  - Waveform and label
  - Speech recognizer
  - Label (sentence)

- **Training SR**
  - Waveform

- **Training WA**
  - Word aligner
  - Label (sentence)

- **Training G2P**
  - Label (word)
  - Full context label
  - G2P converter

- **Speech synthesizer**
  - Full context label
  - Synthesized speech

- **Indian text database**
  - Indian text
  - Input text
  - English text
  - Festival

- **Full context label**
  - Label (sentence)
Speech recognizer

- Initial SR is built by using English
  - WSJ0, WSJ1, and TIMIT databases are used
- SR is built by using recognized label sequences
- To obtain high accuracy SR, SR is re-trained

⇒ Obtain label sequences of target Indian language speech
System overview

- **English database**
- **Indian speech database**
- **Indian text database**

**Training SR**
- **Speech recognizer (initial)**
- **Input text**
  - **Indian text**
  - **English text**

**Training WA**
- **Label (sentence)**

**Training G2P**
- **Word aligner**
- **Label (word)**
- **F2P converter**

**Speech synthesizer**
- **Full context label**
- **Synthesized speech**

**Festival**
- **Full context label**

**Word aligner**
- **Label (word)**

**Training WA**
- **Label (sentence)**

**Training SR**
- **Speech recognizer**
- **Waveform and label**

**Training SS**
- **Speech synthesizer**
- **Waveform**
- **Label (sentence)**
**Word aligner (WA)**

### Word breaking information
- Word breaking information is required for full context labels of speech synthesis
- Word-level G2P converter is required

### Multigram word aligner [Deligne, et al.]
- Multigram models are estimated by using EM algorithm
- Word alignment is obtained by applying Viterbi algorithm

⇒ Obtain word breaking information of label sequences
Speech synthesizer (SS)

Training part

Speech database

Speech signal

Excitation parameters extraction

Spectral parameters extraction

Label

Contest-dependent HMMs & duration models

Training of HMM

Synthesis part

TEXT

Text analysis

Parameter generation from HMM

Excitation generation

Synthesis filter

Synthesized speech

Excitation parameters

Spectral parameters
Base techniques of SS

HSMM [Zen², et al.]
- HMM with explicit state duration probability distribution

MSD [Tokuda², et al.]
- Output distributions consist of continuous dist. and discrete dist.

STRAIGHT [Kawahara, et al.]
- High quality speech vocoding method
- F0, spectrum, and aperiodicity measures

GV [Toda, et al.]
- Intra-utterance variance of speech-parameter trajectory
System overview

English database
Indian speech database

Training SR

Speech recognizer (initial)

Label (sentence)

Label (sentence)

Training SR

Speech recognizer

Waveform and label

Full context label

Training G2P

G2P converter

Indian text

Indian text

Input text

English text

Label (word)

Full context label

Festival

Label (sentence)

Synthesized speech

Speech synthesizer

Speech recognizer

Word aligner

Training WA

Training SS

Training SR

Training SR

Word aligner

Speech synthesizer

Label (sentence)
Grapheme-to-phoneme (G2P)

Joint multigram G2P converter [Bisani, et al.]

- Optimal grapheme and phoneme pair alignment is estimated
- Joint multigram models are estimated by using EM algorithm
- G2P converter is trained by using Sequitur G2P

⇒ Obtain label sequences of input text of target Indian language
Advantage of our system

Multilingual speech synthesis

• Phoneset of acoustic model is the same as the English speech recognizer

• Available text analysis results of the English

• English text analysis: Festival

• Indian language text analysis: G2P converter

Language-independent

• Can apply to languages in which sentences written with a space between words

• e.g. Indian language, Spanish, Arabic
Outline

Background
Blizzard Challenge 2014 rules
System overview
  • Speech recognizer (SR)
  • Word aligner (WA)
  • Speech synthesizer (SS)
  • Grapheme-to-phoneme (G2P) converter

Experiments
Conclusions
<table>
<thead>
<tr>
<th><strong>Speech recognition conditions</strong></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>English database</strong></th>
<th>WSJ0, WSJ1, and TIMIT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Indian database</strong></td>
<td>Six Indian language</td>
</tr>
<tr>
<td><strong>Window</strong></td>
<td>Hamming window</td>
</tr>
<tr>
<td><strong>Frame length</strong></td>
<td>25 ms</td>
</tr>
<tr>
<td><strong>Frame shift</strong></td>
<td>10 ms</td>
</tr>
<tr>
<td><strong>Feature vector</strong></td>
<td>12-dimension MFCC + Δ + ΔΔ (39 dimension)</td>
</tr>
<tr>
<td><strong>HMM</strong></td>
<td>3-state left-to-right HMM without skip transition</td>
</tr>
<tr>
<td><strong>Insertion penalty</strong></td>
<td>−30.0</td>
</tr>
<tr>
<td><strong>Number of iteration</strong> (target language SR)</td>
<td>2</td>
</tr>
</tbody>
</table>
### Speech synthesis conditions

<table>
<thead>
<tr>
<th>Sampling rate</th>
<th>16.0 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window</td>
<td>f0-adaptive Gaussian window</td>
</tr>
<tr>
<td>Frame shift</td>
<td>5 ms</td>
</tr>
<tr>
<td>Feature vector</td>
<td>39-dimension STRAIGHT mel-cepstrum, log f0, 19 aperiodicity measure + Δ + ΔΔ (183 dimension)</td>
</tr>
<tr>
<td>HMM</td>
<td>5-state left-to-right MSD-HSMM without skip transition</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Assamese</th>
<th>Gujarati</th>
<th>Hindi</th>
<th>Rajasthani</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sentences</td>
<td>1427</td>
<td>450</td>
<td>875</td>
<td>1369</td>
<td>822</td>
<td>1470</td>
</tr>
<tr>
<td>Time</td>
<td>2h3m11s</td>
<td>2h1m33s</td>
<td>2h0m31s</td>
<td>2h13m22s</td>
<td>1h57m48s</td>
<td>3h6m32s</td>
</tr>
</tbody>
</table>
Evaluation conditions

<table>
<thead>
<tr>
<th>Evaluation criteria</th>
<th>Intelligibility (WER), similarity (MOS), naturalness (MOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>System A</td>
<td>Natural speech</td>
</tr>
<tr>
<td>System C</td>
<td>NITECH system</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of listeners</th>
<th>Assamese</th>
<th>Gujarati</th>
<th>Hindi</th>
<th>Rajasthani</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>50</td>
<td>109</td>
<td>110</td>
<td>109</td>
<td>106</td>
<td></td>
</tr>
</tbody>
</table>

- SUS: semantically unpredictable sentences
- RD: read text
- ML: multilingual sentences (Indian and English)
## Word error rates (SUS)

<table>
<thead>
<tr>
<th>Language System</th>
<th>Assamese</th>
<th>Gujarati</th>
<th>Hindi</th>
<th>Rajasthani</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>51</td>
<td>24</td>
<td>22</td>
<td>62</td>
<td>32</td>
<td>40</td>
</tr>
<tr>
<td>B</td>
<td>86</td>
<td>34</td>
<td>26</td>
<td>100</td>
<td>33</td>
<td>55</td>
</tr>
<tr>
<td>C</td>
<td><strong>84</strong></td>
<td><strong>59</strong></td>
<td><strong>40</strong></td>
<td><strong>67</strong></td>
<td><strong>64</strong></td>
<td><strong>77</strong></td>
</tr>
<tr>
<td>D</td>
<td>69</td>
<td>40</td>
<td>24</td>
<td>65</td>
<td>38</td>
<td>54</td>
</tr>
<tr>
<td>E</td>
<td>76</td>
<td>23</td>
<td>27</td>
<td>60</td>
<td>37</td>
<td>51</td>
</tr>
<tr>
<td>F</td>
<td>67</td>
<td>25</td>
<td>24</td>
<td>64</td>
<td>37</td>
<td>46</td>
</tr>
<tr>
<td>G</td>
<td>74</td>
<td>37</td>
<td>29</td>
<td>59</td>
<td>37</td>
<td>51</td>
</tr>
<tr>
<td>H</td>
<td>-</td>
<td>41</td>
<td>30</td>
<td>67</td>
<td>60</td>
<td>57</td>
</tr>
<tr>
<td>I</td>
<td>69</td>
<td>44</td>
<td>30</td>
<td>57</td>
<td>34</td>
<td>62</td>
</tr>
<tr>
<td>J</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>44</td>
</tr>
<tr>
<td>K</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
## Similarity and Naturalness (RD)

<table>
<thead>
<tr>
<th>Language System</th>
<th>Assamese</th>
<th>Gujarati</th>
<th>Hindi</th>
<th>Rajasthani</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3.3</td>
<td>4.7</td>
<td>2.9</td>
<td>4.7</td>
<td>4.3</td>
<td>4.5</td>
</tr>
<tr>
<td>B</td>
<td>1.8</td>
<td>2.1</td>
<td>3.0</td>
<td>2.6</td>
<td>2.4</td>
<td>2.0</td>
</tr>
<tr>
<td>C</td>
<td>2.8</td>
<td>3.3</td>
<td>3.0</td>
<td>2.8</td>
<td>2.6</td>
<td>2.5</td>
</tr>
<tr>
<td>D</td>
<td>3.2</td>
<td>3.5</td>
<td>2.7</td>
<td>2.8</td>
<td>4.0</td>
<td>3.6</td>
</tr>
<tr>
<td>E</td>
<td>2.6</td>
<td>2.9</td>
<td>3.5</td>
<td>3.5</td>
<td>3.2</td>
<td>3.1</td>
</tr>
<tr>
<td>F</td>
<td>2.9</td>
<td>3.4</td>
<td>2.8</td>
<td>3.4</td>
<td>3.4</td>
<td>3.2</td>
</tr>
<tr>
<td>G</td>
<td>3.2</td>
<td>3.9</td>
<td>3.7</td>
<td>3.8</td>
<td>3.4</td>
<td>3.7</td>
</tr>
<tr>
<td>H</td>
<td>-</td>
<td>-</td>
<td>3.5</td>
<td>2.5</td>
<td>2.1</td>
<td>2.2</td>
</tr>
<tr>
<td>I</td>
<td>1.8</td>
<td>2.1</td>
<td>2.8</td>
<td>2.7</td>
<td>3.1</td>
<td>2.2</td>
</tr>
<tr>
<td>J</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.1</td>
</tr>
<tr>
<td>K</td>
<td>-</td>
<td>-</td>
<td>2.4</td>
<td>3.4</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Left: MOS of similarity    Right: MOS of naturalness
<table>
<thead>
<tr>
<th>Language System</th>
<th>Language</th>
<th>Assamese</th>
<th>Gujarati</th>
<th>Hindi</th>
<th>Rajasthani</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Assamese</td>
<td>3.8</td>
<td>4.9</td>
<td>3.7</td>
<td>4.7</td>
<td>3.7</td>
<td>4.3</td>
</tr>
<tr>
<td>B</td>
<td>Gujarati</td>
<td>1.6</td>
<td>1.9</td>
<td>2.7</td>
<td>3.0</td>
<td>1.9</td>
<td>2.0</td>
</tr>
<tr>
<td>C</td>
<td>Hindi</td>
<td>2.5</td>
<td>2.8</td>
<td>2.5</td>
<td>2.6</td>
<td>2.7</td>
<td>2.6</td>
</tr>
<tr>
<td>D</td>
<td>Rajasthani</td>
<td>2.8</td>
<td>2.7</td>
<td>2.3</td>
<td>2.5</td>
<td>3.3</td>
<td>2.8</td>
</tr>
<tr>
<td>E</td>
<td>Tamil</td>
<td>2.3</td>
<td>2.2</td>
<td>3.5</td>
<td>2.9</td>
<td>2.5</td>
<td>2.6</td>
</tr>
<tr>
<td>F</td>
<td>Telugu</td>
<td>-</td>
<td>-</td>
<td>1.9</td>
<td>2.8</td>
<td>3.1</td>
<td>3.2</td>
</tr>
<tr>
<td>G</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>H</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></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>
</tr>
<tr>
<td>J</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>2.7</td>
</tr>
<tr>
<td>K</td>
<td>-</td>
<td>-</td>
<td>2.4</td>
<td>3.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Left: MOS of similarity  Right: MOS of naturalness
## Speech samples

<table>
<thead>
<tr>
<th></th>
<th>Assamese</th>
<th>Gujarati</th>
<th>Hindi</th>
<th>Rajasthani</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD</td>
<td><img src="Sound" alt="RD Assamese" /></td>
<td><img src="Sound" alt="RD Gujarati" /></td>
<td><img src="Sound" alt="RD Hindi" /></td>
<td><img src="Sound" alt="RD Rajasthani" /></td>
<td><img src="Sound" alt="RD Tamil" /></td>
<td><img src="Sound" alt="RD Telugu" /></td>
</tr>
<tr>
<td>SUS</td>
<td><img src="Sound" alt="SUS Assamese" /></td>
<td><img src="Sound" alt="SUS Gujarati" /></td>
<td><img src="Sound" alt="SUS Hindi" /></td>
<td><img src="Sound" alt="SUS Rajasthani" /></td>
<td><img src="Sound" alt="SUS Tamil" /></td>
<td><img src="Sound" alt="SUS Telugu" /></td>
</tr>
<tr>
<td>ML</td>
<td><img src="Sound" alt="ML Assamese" /></td>
<td><img src="Sound" alt="ML Gujarati" /></td>
<td><img src="Sound" alt="ML Hindi" /></td>
<td><img src="Sound" alt="ML Rajasthani" /></td>
<td><img src="Sound" alt="ML Tamil" /></td>
<td><img src="Sound" alt="ML Telugu" /></td>
</tr>
</tbody>
</table>

- Generate multilingual speech
- Need to improve intelligibility
Outline

Background
Blizzard Challenge 2014 rules
System overview
  • Speech recognizer (SR)
  • Word aligner (WA)
  • Speech synthesizer (SS)
  • Grapheme-to-phoneme (G2P) converter
Experiments
Conclusions
Conclusions

TTS developed for Blizzard Challenge 2014
• System was built without the phoneme information and phoneset of target Indian language
• Can apply to languages in which sentences written with a space between words
• Generate multilingual speech
• Generate low intelligible speech
  ▪ There is still room for improvement

Future work
• Improve accuracy of G2P converter
• Evaluation in other languages
References


Thank you