The NITECH HMM-based text-to-speech system for the Blizzard Challenge 2015

Kei Sawada, Kei Hashimoto, Keiichiro Oura, and Keiichi Tokuda

Nagoya Institute of Technology (NITECH)

Blizzard Challenge 2015 workshop on Sep 11, 2015
Outline

- Background
  - Blizzard Challenge 2015 tasks
  - Text-to-speech system
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions
Background

- **Text-to-speech (TTS) systems**
  - System to synthesize speech for arbitrary text
  - TTS have been used widely in various applications
    - Car navigation, smart phone, spoken dialogue system, etc.
  - Demand for TTS systems has increased
    - High speech quality, multilingual language, speaking styles, etc.

- **Multilingual language TTS systems**
  - Thousands of languages exist in the world
  - To establish a framework that can be applied to build TTS system of any target languages
    ⇒ One goal of speech synthesis research
  - Require a special knowledge of a target language

Focus on automatically constructing TTS of any languages
Outline

- Background
- Blizzard Challenge 2015 tasks
- Text-to-speech system
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions
Blizzard Challenge 2015 tasks

- Blizzard Challenge [Black, et al.; ’05]
  - Blizzard Challenge was started in order to better understand and compare research techniques

- Speech synthesis for six Indian languages
  - Bengali, Hindi, Malayalam, Marathi, Tamil, Telugu

- Hub task
  - Build TTS systems in each Indian language
  - Provide speech data and corresponding to text

- Spoke task
  - Build a multilingual (polyglot) TTS system (Indian and English)
  - Training data for the Spoke task is same as for the Hub task
  - Sample input text (Hindi and English):
    Stanford वैज्ञानिकों द्वारा विकसित नयी aluminium battery, केवल एक minute में cellphone को charge कर सकती है.
Blizzard Challenge 2015 tasks

Waveform: 

Need to construct TTS system from only speech data and corresponding to text

Text: पृसदिद्ध कबीर अध्येता, पुरुषोत्तम अग्रवाल का यह शोध आलेख, …
Outline

- Background
- Blizzard Challenge 2015 tasks
- **Text-to-speech system**
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions
TTS system components

- **Text analysis part**
  - Estimate pronunciation of an input text
  - Using a lexicon containing pronunciation information
  - **High language-dependency**

- **Speech waveform generation part**
  - Speech waveforms are generated from pronunciation info.
  - Unit-selection [Hunt, et al. ’96]
  - Statistical parametric speech synthesis (SPSS)
    - Hidden Markov model (HMM) [Tokuda, et al.; ’00]
    - Deep neural network (DNN) [Zen, et al.; ’13]
  - **Low language-dependency**
TTS system construction

- Steps of TTS system construction
  - Acoustic features are modeled by using phoneme information

  Target language DB
  Speech data
  Text

  बहुत धन्यवाद

  - Require a special knowledge of the target language
    ⇒ Huge cost for someone not familiar with the target language

Step 1: Definition of a phoneset
Step 2: Construction of a lexicon or G2P
Step 3: Design of contextual factors
Step 4: Preparation of label sequences

TTS system

Investigate a framework to automatic TTS construction in any unknown-pronunciation languages
Outline

- Background
- Blizzard Challenge 2015 tasks
- Text-to-speech system

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

- Experiments
- Conclusions
System overview

Target language text database

English speech and label database

Target language speech database

Text (w/ WB)

Waveform and label

Label (w/o WB)

Training SR

SR

Label (w/ WB)

Training WA

WA

Full-context label

Training G2P

G2P

Input text (w/ WB)

Training SS

SS

Synthesized speech

Database

Process

Component

SR: Speech recognizer
WA: Word aligner
G2P: Grapheme-to-phoneme converter
SS: Speech synthesizer
WB: Word boundary

Training part

Synthesis part
System overview

Text (w/ WB) → Target language text database (बहुत धन्यवाद)

→ English speech and label database

→ Waveform and label

Training SR → SR

- Label (w/o WB)

→ Training WA

- Full-context label

→ Training G2P

- Full-context label

Input text (w/ WB) → G2P

Training SS → SS

- Synthesized speech

→ Target language speech database

→ Waveform

SR: Speech recognizer
WA: Word aligner
G2P: Grapheme-to-phoneme converter
SS: Speech synthesizer
WB: Word boundary

: Database
: Process
: Component

Training part

Synthesis part
Speech recognizer (SR) (1/3)

- Phoneme sequence corresponding to speech data
  - Speech recognition is carried out by using speaker-independent SR (SISR) of the other language, e.g., English
  - Triphone recognizer
  - Phoneset of SISR is used as phonset of target language
Construction flow

Waveform:

SISR output:  sil th ih s ah t uh g ah b iy uh ih hh ih k ah sil …

Text:  पृसदिद कबीर अध्येता, पुरुषोत्तम अग्रवाल का यह शोध आलेख, …
Speech recognizer (SR) (2/3)

- More accurate phoneme sequence
  - Speaker-dependent SR (SDSR) is constructed from phoneme sequences obtained by the SISR.
  - Estimation of phoneme sequences and training of SDSR are iterated

- Unsupervised training
  - Phoneme sequences obtained by the SISR is initial value
  - Data-driven training by repetition

![Diagram showing the process of training SISR and SDSR]
Construction flow

Waveform:

SISR output: sil th ih s ah t uh g ah b iy uh ih hh ih k ah sil …

SDSR1 output: sil ah s uh b t ah g ah b iy d ah d hh ih t ae sil …

SDSR2 output: sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …

Text: पृषद्दिद कबीर अध्येता, पुरुषोत्तम अग्रवाल का यह शोध आलेख, …
Speech recognizer (SR) (3/3)

- Phoneme sequence taking account of phoneme duration
  - Modeling of phoneme durations is important for an SS
  - SR is difficult to consider phoneme duration
- Phoneme sequence is selected using alignment likelihood of HSMM
- Phoneme sequence with the highest alignment likelihood is selected as phoneme sequence
Construction flow

Waveform:

SISR output: sil th ih s ah t uh g ah b iy uh ih hh ih k ah sil …

SDSR1 output: sil ah s uh b t ah g ah b iy d ah d hh ih t ae sil …

SDSR2 output: [sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …]

N-best [sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …]

HSMM output: sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …

Obtain phoneme sequences of target language speech
System overview

Text (w/ WB) → Target language text database

Input text (w/ WB) → G2P → Training G2P → Full-context label → Synthesized speech → SS → Target language speech database

Training part

SR: Speech recognizer
WA: Word aligner
G2P: Grapheme-to-phoneme converter
SS: Speech synthesizer
WB: Word boundary

Synthesis part

Label (w/o WB) → Training WA → WA → Full-context label → Label (w/ WB) → Training SR → SR → Waveform and label → English speech and label database

Database : Database
Process : Process
Component : Component
Word aligner (WA)

- **Word boundary (WB) information**
  - Word-level G2P is suitable
  - WB information is useful for contextual factors of the SS
  - Phoneme sequences obtained by speech recognition does not include WB

- **Joint multigram model-based WA**
  - WA is constructed for estimation of WB
  - A pause of recognition results must be WB
  - Viterbi decoding

1. Target lang. text DB
   - बहुत धन्यवाद
2. Text (w/ WB) बहुत धन्यवाद
3. Training WA
4. WA
5. Label (w/ WB)
   - a g c d f …
6. Label (w/o WB)
   - a g c | d f …
Construction flow

Waveform:

SISR output: sil th ih s ah t uh g ah b iy uh ih hh ih k ah sil …

SDSR1 output: sil ah s uh b t ah g ah b iy d ah d hh ih t ae sil …

SDSR2 output: sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …

N-best

Obtain WB information of phoneme sequence

HSMM output:

WA output: sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …

Text:

प्रसदिद व कबीर अध्येता, ...

Training WA

WA
System overview

- **Target language text database**
- **English speech and label database**
- **Target language speech database**

**Training part**
- Text (w/ WB)
- Input text (w/ WB)
- Label (w/ WB)
- Training G2P
- Full-context label

**Process**
- SR: Speech recognizer
- WA: Word aligner
- G2P: Grapheme-to-phoneme converter
- SS: Speech synthesizer
- WB: Word boundary

**Component**
- Database
- Process
- Component

**Synthesis part**
- Synthesized speech
Grapheme-to-phoneme converter (G2P) (1/2)

- Arbitrary input text need to be converted into phoneme sequence
  - Difficult to construct a lexicon in unknown-pronunciation language

- Joint multigram model-based G2P [Bisani, et al.; '08]
  - G2P is introduced instead of lexicon
  - Viterbi decoding

- Pause insertion
  - Comma, colon, parenthesis
  - Before or after a word that is easy to enter pause in a speech recognition result
Construction flow

Waveform:

SISR output: sil th ih s ah t uh g ah b iy uh ih hh ih k ah sil …

SDSR1 output: sil ah s uh b t ah g ah b iy d ah d hh ih t ae sil …

SDSR2 output: sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …
N-best

Obtain phoneme sequence of arbitrary input text

HSMM output:

WA output: sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …

Text: प्रसिद्ध कबीर अध्येता, ...

Training G2P

G2P
Multilingual speech synthesis
- Input text includes Indian language and English
- Phoneset of acoustic model is the same as the English SISR
- Can synthesize Indian language and English
- Indian language text analysis: G2P
- English text analysis: Festival

Can synthesize multilingual speech
System overview

- Target language text database
- English speech and label database
  - Waveform and label
    - Training SR
      - SR
        - Label (w/o WB)
          - Training WA
            - WA
              - Full-context label
                - Training G2P
                  - G2P
                    - Full-context label
                      - Input text (w/ WB)

- Target language speech database
  - Waveform
    - Full-context label
      - Training SS
        - SS
          - Synthesized speech

Text (w/ WB) -> Waveform and label -> Training SR -> SR -> Label (w/o WB) -> Training WA -> WA -> Full-context label -> Training G2P -> G2P -> Full-context label -> Synthesis part

Training part

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

- Design of contextual factors
  - Quint phone
  - Syllable
    - Defined as $C^*V$ ($C$: consonant, $V$: vowel, $C^*$: none or more $C$)
  - Word
  - Phrase
  - Utterance

- Base techniques
Construction flow

Waveform:

SISR output:  sil th ih s ah t uh g ah b iy uh ih hh ih k ah sil …

SDSR1 output:  sil ah s uh b t ah g ah b iy d ah d hh ih t ae sil …

SDSR2 output:  sil r ah s uw b r uh g ah b iy d ah d hh ih t ae sil …
  N-best  sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …
  :  :

HSMM output:  sil r ah s ih d r uh g ah b iy d ah d hh ih t ae sil …

WA output:  sil r ah s ih d r uh | g ah b iy d | ah d hh ih t ae | sil …

Can synthesize speech from full-context label
Strengths and weaknesses

- **Strengths**
  - Low language-dependency
    - Can apply to languages in which sentences written with space between words
  - Multilingual speech synthesis
    - Can synthesize target language and English

- **Weaknesses**
  - Pronunciation (text analysis) error
    - Pronunciation errors can occur due to errors in SR, WA, and G2P
  - Difficult to adjust each component
    - I don’t understand what synthesized speech is saying
Outline

- Background
- Blizzard Challenge 2015 tasks
- Text-to-speech system
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions
# Speech recognizer conditions

<table>
<thead>
<tr>
<th>English database</th>
<th>WSJ0, WSJ1, and TIMIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indian database</td>
<td>Six Indian language</td>
</tr>
<tr>
<td>Window</td>
<td>Hamming window</td>
</tr>
<tr>
<td>Frame</td>
<td>length: 25 ms, shift: 10 m</td>
</tr>
<tr>
<td>Feature vector</td>
<td>12-dimension MFCC + Δ + ΔΔ</td>
</tr>
<tr>
<td>Model structure</td>
<td>3-state left-to-right HMM without skip transition</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bengali</th>
<th>Hindi</th>
<th>Malayalam</th>
<th>Marathi</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertion penalty</td>
<td>-20</td>
<td>-40</td>
<td>-40</td>
<td>-40</td>
<td>-20</td>
</tr>
<tr>
<td>Number of iteration</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Speech synthesizer conditions

<table>
<thead>
<tr>
<th>Indian database</th>
<th>Six Indian language</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sampling rate</strong></td>
<td><strong>16.0 kHz</strong></td>
</tr>
<tr>
<td><strong>Window</strong></td>
<td><strong>F0-adaptive Gaussian window</strong></td>
</tr>
<tr>
<td><strong>Frame shift</strong></td>
<td><strong>5 ms</strong></td>
</tr>
<tr>
<td><strong>Feature vector</strong></td>
<td><strong>39-dim. STRAIGHT mel-cepstrum, log F0, 19-dmi. aperiodicity measure + Δ + ΔΔ</strong></td>
</tr>
<tr>
<td><strong>Model structure</strong></td>
<td><strong>5-state left-to-right MSD-HSMM without skip transition</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Bengali</th>
<th>Hindi</th>
<th>Malayalam</th>
<th>Marathi</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of sentences</strong></td>
<td>1284</td>
<td>1690</td>
<td>1269</td>
<td>1178</td>
<td>1440</td>
<td>2461</td>
</tr>
<tr>
<td><strong>Time</strong></td>
<td>1h58m27s</td>
<td>3h57m59s</td>
<td>1h58m9s</td>
<td>2h6m7s</td>
<td>4h9m28s</td>
<td>4h11m34s</td>
</tr>
</tbody>
</table>
# Evaluation conditions

<table>
<thead>
<tr>
<th>Evaluation criteria</th>
<th>Intelligibility (word error rate), MOS of speaker similarity, MOS of naturalness</th>
</tr>
</thead>
<tbody>
<tr>
<td>System A</td>
<td>Natural speech</td>
</tr>
<tr>
<td>System G</td>
<td>NITECH system</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RD</th>
<th>Read text</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUS</td>
<td>Semantically unpredictable sentences</td>
</tr>
<tr>
<td>ML</td>
<td>Multilingual sentences (Indian and English)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of listeners</th>
<th>Bengali</th>
<th>Hindi</th>
<th>Malayalam</th>
<th>Marathi</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>69</td>
<td>72</td>
<td>69</td>
<td>70</td>
<td>70</td>
<td></td>
</tr>
</tbody>
</table>
### Word error rate (SUS)

<table>
<thead>
<tr>
<th></th>
<th>Bengali</th>
<th>Hindi</th>
<th>Malayalam</th>
<th>Marathi</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WER</td>
<td>WER</td>
<td>WER</td>
<td>WER</td>
<td>WER</td>
<td>WER</td>
</tr>
<tr>
<td>A</td>
<td>43</td>
<td>42</td>
<td>59</td>
<td>35</td>
<td>29</td>
<td>82</td>
</tr>
<tr>
<td>B</td>
<td>52</td>
<td>49</td>
<td>64</td>
<td>35</td>
<td>57</td>
<td>70</td>
</tr>
<tr>
<td>C</td>
<td>100</td>
<td>34</td>
<td>50</td>
<td>26</td>
<td>49</td>
<td>46</td>
</tr>
<tr>
<td>D</td>
<td>57</td>
<td>25</td>
<td>78</td>
<td>30</td>
<td>49</td>
<td>60</td>
</tr>
<tr>
<td>E</td>
<td>69</td>
<td>40</td>
<td>47</td>
<td>52</td>
<td>69</td>
<td>54</td>
</tr>
<tr>
<td>F</td>
<td>66</td>
<td>23</td>
<td>74</td>
<td>41</td>
<td>67</td>
<td>59</td>
</tr>
<tr>
<td>G</td>
<td>76</td>
<td>30</td>
<td>98</td>
<td>69</td>
<td>47</td>
<td>75</td>
</tr>
<tr>
<td>H</td>
<td>55</td>
<td>24</td>
<td>46</td>
<td>35</td>
<td>57</td>
<td>46</td>
</tr>
<tr>
<td>I</td>
<td>100</td>
<td>31</td>
<td>73</td>
<td>-</td>
<td>50</td>
<td>62</td>
</tr>
<tr>
<td>J</td>
<td>61</td>
<td>40</td>
<td>52</td>
<td>21</td>
<td>60</td>
<td>57</td>
</tr>
</tbody>
</table>

**Important to properly adjust SR, WA, G2P for each language**
MOS of speaker similarity

<table>
<thead>
<tr>
<th></th>
<th>Bengali</th>
<th></th>
<th></th>
<th>Hindi</th>
<th></th>
<th></th>
<th>Malayalam</th>
<th></th>
<th></th>
<th>Marathi</th>
<th></th>
<th></th>
<th>Tamil</th>
<th></th>
<th></th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RD</td>
<td>SUS</td>
<td>ML</td>
<td>RD</td>
<td>SUS</td>
<td>ML</td>
<td>RD</td>
<td>SUS</td>
<td>ML</td>
<td>RD</td>
<td>SUS</td>
<td>ML</td>
<td>RD</td>
<td>SUS</td>
<td>ML</td>
<td>RD</td>
</tr>
<tr>
<td>A</td>
<td>4.5</td>
<td>4.7</td>
<td>4.6</td>
<td>4.5</td>
<td>4.4</td>
<td>4.5</td>
<td>4.6</td>
<td>4.2</td>
<td>4.2</td>
<td>4.4</td>
<td>4.3</td>
<td>4.3</td>
<td>4.6</td>
<td>4.6</td>
<td>4.2</td>
<td>4.5</td>
</tr>
<tr>
<td>B</td>
<td>2.5</td>
<td>3.7</td>
<td>1.8</td>
<td>2.6</td>
<td>3.5</td>
<td>1.8</td>
<td>1.8</td>
<td>2.1</td>
<td>2.6</td>
<td>2.3</td>
<td>2.7</td>
<td>2.2</td>
<td>1.8</td>
<td>1.9</td>
<td>2.2</td>
<td>2.1</td>
</tr>
<tr>
<td>C</td>
<td>2.2</td>
<td>2.2</td>
<td>2.1</td>
<td>2.4</td>
<td>2.0</td>
<td>2.2</td>
<td>2.3</td>
<td>2.1</td>
<td>2.2</td>
<td>2.3</td>
<td>1.9</td>
<td>1.6</td>
<td>2.2</td>
<td>2.8</td>
<td>2.8</td>
<td>1.3</td>
</tr>
<tr>
<td>D</td>
<td>2.2</td>
<td>3.0</td>
<td>2.8</td>
<td>2.7</td>
<td>2.2</td>
<td>2.0</td>
<td>2.1</td>
<td>2.2</td>
<td>2.1</td>
<td>3.0</td>
<td>3.4</td>
<td>3.1</td>
<td>1.9</td>
<td>2.0</td>
<td>1.7</td>
<td>2.0</td>
</tr>
<tr>
<td>E</td>
<td>3.7</td>
<td>3.3</td>
<td>4.1</td>
<td>2.9</td>
<td>2.8</td>
<td>3.5</td>
<td>2.9</td>
<td>2.6</td>
<td>3.2</td>
<td>2.9</td>
<td>3.4</td>
<td>2.7</td>
<td>2.7</td>
<td>2.6</td>
<td>3.3</td>
<td>2.9</td>
</tr>
<tr>
<td>F</td>
<td>1.7</td>
<td>2.6</td>
<td>2.1</td>
<td>4.3</td>
<td>3.9</td>
<td>3.1</td>
<td>2.3</td>
<td>2.3</td>
<td>2.6</td>
<td>3.0</td>
<td>2.8</td>
<td>2.7</td>
<td>2.7</td>
<td>2.5</td>
<td>2.6</td>
<td>2.5</td>
</tr>
<tr>
<td>G</td>
<td>3.1</td>
<td>3.1</td>
<td>2.2</td>
<td>2.8</td>
<td>2.9</td>
<td>2.2</td>
<td>2.3</td>
<td>2.1</td>
<td>2.0</td>
<td>2.5</td>
<td>2.2</td>
<td>2.1</td>
<td>2.3</td>
<td>3.1</td>
<td>2.4</td>
<td>3.1</td>
</tr>
<tr>
<td>H</td>
<td>2.4</td>
<td>2.5</td>
<td>-</td>
<td>2.7</td>
<td>2.6</td>
<td>-</td>
<td>2.0</td>
<td>2.2</td>
<td>-</td>
<td>2.1</td>
<td>2.4</td>
<td>-</td>
<td>2.6</td>
<td>2.7</td>
<td>-</td>
<td>2.4</td>
</tr>
<tr>
<td>I</td>
<td>2.9</td>
<td>2.3</td>
<td>-</td>
<td>3.5</td>
<td>3.3</td>
<td>-</td>
<td>3.0</td>
<td>3.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.6</td>
<td>3.4</td>
<td>-</td>
<td>4.2</td>
</tr>
<tr>
<td>J</td>
<td>2.7</td>
<td>2.8</td>
<td>-</td>
<td>3.3</td>
<td>2.2</td>
<td>-</td>
<td>2.0</td>
<td>2.9</td>
<td>-</td>
<td>2.8</td>
<td>2.4</td>
<td>-</td>
<td>2.6</td>
<td>2.3</td>
<td>-</td>
<td>2.7</td>
</tr>
</tbody>
</table>

The high MOS even though WER is the high rate (76%) ⇒ High WER or pronunciation errors scarcely affect MOS
## MOS of naturalness

<table>
<thead>
<tr>
<th></th>
<th>Bengali</th>
<th>Hindi</th>
<th>Malayalam</th>
<th>Marathi</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>RD 4.7</td>
<td>SUS 4.6</td>
<td>ML 4.7</td>
<td>RD 4.3</td>
<td>SUS 4.4</td>
<td>ML 4.4</td>
</tr>
<tr>
<td>B</td>
<td>RD 2.2</td>
<td>SUS 2.7</td>
<td>ML 1.8</td>
<td>RD 1.6</td>
<td>SUS 1.9</td>
<td>ML 1.9</td>
</tr>
<tr>
<td>C</td>
<td>RD 2.9</td>
<td>SUS 3.1</td>
<td>ML 2.6</td>
<td>RD 2.6</td>
<td>SUS 2.8</td>
<td>ML 2.4</td>
</tr>
<tr>
<td>D</td>
<td>RD 3.0</td>
<td>SUS 3.0</td>
<td>ML 2.6</td>
<td>RD 2.8</td>
<td>SUS 2.7</td>
<td>ML 2.2</td>
</tr>
<tr>
<td>E</td>
<td>RD 3.4</td>
<td>SUS 2.7</td>
<td>ML 3.8</td>
<td>RD 2.6</td>
<td>SUS 3.0</td>
<td>ML 3.0</td>
</tr>
<tr>
<td>F</td>
<td>RD 2.0</td>
<td>SUS 1.8</td>
<td>ML 1.6</td>
<td>RD 3.9</td>
<td>SUS 2.9</td>
<td>ML 3.9</td>
</tr>
<tr>
<td>G</td>
<td>RD 2.5</td>
<td>SUS 2.8</td>
<td>ML 2.2</td>
<td>RD 2.4</td>
<td>SUS 2.0</td>
<td>ML 2.1</td>
</tr>
<tr>
<td>H</td>
<td>RD 2.6</td>
<td>SUS 2.6</td>
<td>ML 2.5</td>
<td>RD 2.7</td>
<td>SUS 2.2</td>
<td>ML 3.0</td>
</tr>
<tr>
<td>I</td>
<td>RD 2.7</td>
<td>SUS 2.1</td>
<td>ML 3.2</td>
<td>RD 3.2</td>
<td>SUS 3.0</td>
<td>ML 3.0</td>
</tr>
<tr>
<td>J</td>
<td>RD 2.6</td>
<td>SUS 2.5</td>
<td>ML 3.1</td>
<td>RD 2.7</td>
<td>SUS 2.9</td>
<td>ML 3.5</td>
</tr>
</tbody>
</table>

The low MOS even though WER is the lowest rate (47%) ⇒ Even a little word pronunciation error often affects MOS.
## Speech samples

<table>
<thead>
<tr>
<th></th>
<th>Bengali</th>
<th>Hindi</th>
<th>Malayalam</th>
<th>Marathi</th>
<th>Tamil</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
</tr>
<tr>
<td>SUS</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
</tr>
<tr>
<td>ML</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
</tr>
</tbody>
</table>

Public private partnership model - এমবি মেডিকেল কলেজ

Release of Hollywood film, Fast and Furious Seven, also created a stir in India.

Traffic police officers have received many complaints.

Facebook users are consistently online for their mood or current status updates.

India is my country, இந்தியா என் நாடாண்டு, I am an Indian, என்னும் இலக்கணங்கள்.
Outline

- Background
- Blizzard Challenge 2015 tasks
- Text-to-speech system
- System overview
  - Speech recognizer (SR)
  - Word aligner (WA)
  - Grapheme-to-phoneme converter (G2P)
  - Speech synthesizer (SS)
- Experiments
- Conclusions
Conclusions

- TTS system developed for the Blizzard Challenge 2015
  - Investigated automatic TTS system construction in an unknown-pronunciation language
  - Enabled a target language and multilingual TTS construction
  - Achieved a high score if the SR, WA, and G2P component can properly construct

- Future work
  - Investigation of a construction criteria
  - Construction of the multilingual SISR using the IPA
  - Investigation of phoneset determination approaches based on speech data