Building Speech Synthesis Systems for Indian Languages

Hema A Murthy
hema@cse.iitm.ac.in

TTS Consortium, Funded by DeitY, GoI, India

03 July 2017
Outline

1. Text to Speech Synthesis
2. Languages of India
3. Data Collection
4. Segmentation
5. Syllables and group delay functions
6. DNN/CNNs and boundary correction
   - Experiments and Results
   - Architectures - Sub-utterance level
   - Experiments and Results
7. Scope for Improvement
Given an input text in a particular language, the objective of a TTS system is to produce natural and intelligible speech output.

Applications

- Readers for the visually challenged.
- Readers for small form factor smart phones.
- Enablers for the language challenged.

Objective: Build TTS systems for a number of Indian languages – with a short turnaround time for any new Indian language.
What does speech synthesis involve?

- Collection of speech data with correct transcriptions.
- The text must consist of examples of all sub-word units in various contexts.
- The speech needs to be segmented into sub-word units.
- The sub-words (or their models) are stored in a database.
- During synthesis
  - A sentence is split into subword units.
  - Waveforms (as in Unit Selection Synthesis) or models (as in Statistical Parametric Synthesis) of sub-words (based on context) are concatenated to generate the waveform.
Convergence and divergence of Indian languages

- Indian scripts: based on the ancient Brahmi script.
- Writing system corresponds to Aksharas.
- Indian scripts are syllabic \((C^*VC^*)\) in nature.
- Aksharas: V, CV, CCV, CCCV.
- Aryan and Dravidian – originally less similar have become more similar.
- Basic set of phones: 50; 35-38 consonants and 15-18 vowels.
- Differences in these languages mostly due to phonotactics – not dialectal variations.
- Although some languages can even be characterised by a common syllable set, the prosody – duration, prominence associated with each syllable in a word can be significantly different.
13 Indian languages

Text for optimal text selection: trisyllabic words at best, ≈ 60-70 thousand unique words.

Source: Online newspapers, blogs, children’s stories, avoid proper names.

Pronunciation dictionaries – UTF-8 based – syllable-level 100,000 words.

5 hours of data for each language (male, female, L2 English)
Summary of data collected

Table: Statistics of the total collected text corpus for 13 Indian Languages before and after text optimization

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N.sent</td>
<td>N.wrd</td>
</tr>
<tr>
<td>Hin</td>
<td>55000</td>
<td>582512</td>
</tr>
<tr>
<td>Tam</td>
<td>75000</td>
<td>786548</td>
</tr>
<tr>
<td>Mar</td>
<td>1519950</td>
<td>16419046</td>
</tr>
<tr>
<td>Ben</td>
<td>50000</td>
<td>538124</td>
</tr>
<tr>
<td>Mal</td>
<td>378654</td>
<td>3560283</td>
</tr>
<tr>
<td>Tel</td>
<td>4643</td>
<td>110241</td>
</tr>
<tr>
<td>Kan</td>
<td>41037</td>
<td>365399</td>
</tr>
<tr>
<td>Guj</td>
<td>17000</td>
<td>202847</td>
</tr>
<tr>
<td>Raj</td>
<td>6926</td>
<td>67758</td>
</tr>
<tr>
<td>Ass</td>
<td>1806</td>
<td>32721</td>
</tr>
<tr>
<td>Man</td>
<td>2007</td>
<td>26028</td>
</tr>
<tr>
<td>Odi</td>
<td>35404</td>
<td>737654</td>
</tr>
<tr>
<td>Bod</td>
<td>15000</td>
<td>162400</td>
</tr>
</tbody>
</table>
Figure: Segmentation at syllable and phone level using HMM based flat start

Essentially these segments are used in building models of phones for speech synthesis.

Question: Is it possible to improve the segmentation with small amounts of data?
Segmentation in Speech Systems

- Segmentation is important for both speech synthesis and recognition.
- **Speech recognition**
  - An utterance is split at the word level, which in turn is split into subwords (phone/syllable/triphone).
  - A sequence of words is obtained.
- **Speech synthesis**
  - A given text is split into a sequence of subword units.
  - The waveforms or models are concatenated to generate the waveform.
  - Unlike speech recognition, the consumers of synthesis are the human ears.
  - Accurate segmentation is crucial for speech synthesis.
An Aside: Parsers for Indian Languages
A uniform parser for Indian languages

- A common label set for all Indian languages
- A common set of rules
- Exception handling for each language – a set of special rules
  - taajamahala – taajmahal/taajamhal/taajamahala – depending upon the language and context.
- A lex and yacc based parser is developed for Indian languages – supports 13 Indian languages – 4 Dravidian (Tamil, Telugu, Kannada, Malayalam), 8 Aryan (Hindi, Bengali, Gujarati, Marathi, Rajasthani, Odia, Assamese, Manipuri), and 1 Sino-Tibetan (Bodo).
### Mapping to Common Label Set

<table>
<thead>
<tr>
<th>Hindi Vowels</th>
<th>Hindi Consonants</th>
<th>Hindi Semi Vowels</th>
<th>Hindi Fricatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>आ इ ई उ ऊ ए ऐ ओ औ</td>
<td>क ख ग घ च छ ज झ ञ</td>
<td>य र ल व</td>
<td>त थ द ध न प फ ब व</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tamil Vowels</th>
<th>Tamil Consonants</th>
<th>Tamil Semi Vowels</th>
<th>Fricatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>a aa ax i ii</td>
<td>k kh q gh ng</td>
<td>y r l lx w</td>
<td>sh sx s h</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bengali Vowels</th>
<th>Bengali Consonants</th>
<th>Semi Vowels</th>
<th>Fricatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>a e i u</td>
<td>k kh q gh ng</td>
<td>y r l</td>
<td>sh sx s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Malayalam Vowels</th>
<th>Malayalam Consonants</th>
<th>Semi Vowels</th>
<th>Fricatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>a e i u</td>
<td>k kh q gh ng</td>
<td>y r l</td>
<td>sh sx s</td>
</tr>
</tbody>
</table>
Proposed rules

Hindi words: अकबर असफल follows v-cv-cv-cv structure
But अकबर (a-k-a-b-a-r-a) parsed as a-k-b-a-r (vc-cvc)
and असफल (a-s-a-f-a-l-a) parsed as a-s-a-f-a-l (v-cv-cvc)

Hindi words: ताजमहल पागलपन follows cv-cv-cv-cv-cv structure
But ताजमहल (t-aa-j-a-m-a-h-a-l-a) parsed to t-aa-j-m-a-h-a-l (cvc-cv-cvc)
and पागलपन (p-aa-g-a-l-a-p-a-n-a) parsed to p-aa-g-a-l-p-a-n (cv-cv-cv-cvc)

---

1 Arun Baby N L Nishanthi, Anju L Thomas and Hema A. Murthy, “A Unified parser for developing Indian language text to speech synthesizers”, in International Conference on Text, Speech and Dialogue (TSD), Sept 2016, pp. 514-521.
Agglutination is the process of combining words that are formed by stringing together morphemes.

Unified parser handles the agglutinative words that are common in Dravidian languages since it employs a rule-based approach.

**Tamil**

婆்துகைக்கைத்திருக்கு → பேர் நாலாம் இறக்குமான்

\[ w-a-n-d-u-k-o-nx-dx-i-r-u-k-k-i-rx-aa-n \rightarrow w-a-n-d-u \ k-o-nx-dx-u \ i-r-u-k-k-i-rx-aa-n \]

**Malayalam**

പാന്നിപ്പെട്ടിറക്കുപ്പു → പെരു കണ്ണാര് മതിക്കുപ്പു

\[ w-a-n-n-u-k-o-nx-tx-i-r-i-k-k-u-n-n-u \rightarrow w-a-n-n-u \ k-o-nx-tx \ i-r-i-k-k-u-n-n-u \]
Building speech synthesis systems for Indian languages
Bootstrap Segmentation

- Small amount of data labeled manually at the phone level.
- Phone models are built.
- Forced Viterbi alignment is performed on the rest of the data using the models built.
- Models are rebuilt using the forced aligned data.

Issues: Inconsistencies

- Perceiving phones based on listening and spectrogram is difficult in isolation.
- Inconsistency across annotators.
Flatstart Segmentation

- HMM models initialised such that state means and variances are equal to the global mean and variance.
- Embedded training\(^2\) is performed to build models:
  - Uses transcription to obtain composite HMM for each utterance by concatenating phone HMMs.
  - Embedded Baum-Welch re-estimation.
- Using these models, forced Viterbi alignment is performed to obtain segmentation at phone level.

Drawbacks of HMM Based Segmentation

- A fundamental drawback of this approach is that boundaries are not explicitly modeled \(^3\)
- HMMs do not use proximity to boundary positions as a criterion for optimality during training \(^4\)

---


Example of segmentation at syllable-level is FS HMM (at phone-level)

Figure: Example

Syllable /p a tx/ 🎵

Observe the boundary for the stop consonants at either end is wrong as evidenced by the spectrogram.
Support Vector Machine (SVM) classifiers are used to locate boundaries in Hung-Yi Lo et al., “Phonetic boundary refinement using support vector machines,” ICASSP, 2007. A special one-state HMM is used for detecting phoneme boundaries in 5.

In 6 a multi layer perceptron is used to refine phone boundaries.

---


Drawbacks

- Supervised.
- Manually marked accurate boundaries are required for training.
Can we exploit time domain and spectral cues in the signal to rectify boundaries?
Figure: Inverse of energy as a cue
Figure: Spectral change as a cue
Property of a syllable

Three components: Onset, Rime (nucleus) and Coda

Rime is a sonorant, characterised by a vowel. Syllable definition: (C*VC*)
Energy high in the middle and tapers off towards the end.
Even English has syllables of CV type (70%) – $^7$

$^7$Steven Greenberg, Speech Communication
Group delay functions

Consider a discrete time domain signal \( x[n] \) and its Fourier transform
\[ X(\omega) = |X(\omega)|e^{j\phi(\omega)} \]

\[ \tau(\omega) = -\frac{d\phi(\omega)}{d\omega} \] (1)

Alternate form of computation:
\[ \tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \] (2)

\( X_R, X_I \): Real, Imaginary parts of \( X(\omega) = FFT(x[n]) \)
\( Y_R, Y_I \): Real, Imaginary parts of \( Y(\omega) = FFT(nx[n]) \)

Figure: a) Wrapped, and b) Unwrapped phase response of an Elliptic low pass filter

---

Figure: Resolving power of the group delay spectrum: z-plane, magnitude spectrum and group delay spectrum I) a pole inside the unit circle at $(0.8, \pi/8)$, II) a pole inside the unit circle at $(0.8, \pi/4)$ and III) a pole at $(0.8, \pi/8)$ and another pole at $(0.8, \pi/4)$, inside the unit circle.

Both poles inside the unit circle $\implies$ minimum phase system.
Mixed phase system behaviour

Figure: Group delay property of different types of signals: minimum and nonminimum phase signals

Clearly, system is best behaved for minimum phase system.
Revisiting the source system model for speech

**Figure:** A source system model for speech production

- System: stable and causal $\implies$ poles inside unit circle
- System: zeroes may lie inside or outside – nasals
- System: No zeroes on unit circle – even zeroes have finite bandwidth
Feature extraction from phase: Zeroes on the unit circle

Figure: Significance of proximity of zeros to the unit circle
Group delay processing

- Linear prediction based group delay spectra.
- Root cepstrum based group delay spectra.
- Modified group delay spectra.
Short-term Energy as an arbitrary magnitude spectrum

- Energy is a positive function
- Symmetrise energy
- Arbitrary magnitude function
- Minimum phase group delay function
- Valleys correspond to syllable boundaries approximately
Segmentation of speech using group delay functions

Speech signal → Short term energy computation \( E(m) \) → Invert STE and raise the power by "gamma" \( 0 < \text{gamma} < 1 \) → Symmetrise

Syllable boundaries → Detection of Peaks → Group delay spectrum → Single sided Hanning Window & Group delay processing → Root cepstrum → Compute IDFT
Enforcing syllable boundaries during Embedded Re-estimation\textsuperscript{10}

- Baum-Welch embedded re-estimation\textsuperscript{9} is performed at the syllable level to build phone HMM models.
- Forced alignment also performed within the syllable

\textsuperscript{9}S. Young et al., The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2002.

\textsuperscript{10}S Aswin Shanmugam and Hema A Murthy, Group Delay Based Phone Segmentation for HTS, in Proc. of Twentieth National Conference on Communications (NCC 2014)
### Example

<table>
<thead>
<tr>
<th>Syllable Label</th>
<th>Phone Label</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Beg</strong></td>
<td><strong>End</strong></td>
</tr>
<tr>
<td>0.000</td>
<td>0.234</td>
</tr>
<tr>
<td>0.234</td>
<td>0.363</td>
</tr>
<tr>
<td>0.363</td>
<td>0.516</td>
</tr>
<tr>
<td>0.516</td>
<td>0.647</td>
</tr>
<tr>
<td>0.647</td>
<td>0.728</td>
</tr>
<tr>
<td>0.728</td>
<td>1.113</td>
</tr>
<tr>
<td>1.113</td>
<td>1.228</td>
</tr>
</tbody>
</table>

**Table:** Phone labels from syllable labels
Figure: Example – without boundary correction

Syllable /p a tx/
Segmenation Correction

**Figure:** Example – with boundary correction

Syllable /p a tx/
Spectral Flux based cues

- Spectral Flux is the Euclidean distance between the normalised power spectrum of one frame and the normalised power spectrum of the previous frame
Figure: Sibilant Fricatives and Affricates Boundary Detection
Group Delay Functions

Speech signal

Sub-Band Spectral Flux (SBSF) computation

SBSF

Symmetrise SBSF function

E[k], Magnitude spectrum of an arbitrary signal

Raise the power of E[k] by "gamma"

0 < gamma < 1

Magnitude spectrum

Remove mean, normalize, detect peaks and apply threshold

Single sided Hanning window and compute magnitude spectrum

Root cepstrum

Compute IDFT

Figure: Algorithm
Hybrid Segmentation

- Uses both Short-Time Energy (STE) and Sub-Band Spectral Flux (SBSF).
- While syllable level reestimation add context to phones (eg. “d uu k” becomes “beg-d uu k_end”).
- Correction rules are different.
- Uses threshold
Boundary Correction Rules

- syllable1 boundary syllable2
- $P^* e_p$ boundary $b_p P^*$
- If $b_p == \text{(Unvoiced Stop Consonant)}$, use STE with $threshold_1$ for correction.
- If $e_p == \text{(Unvoiced Stop Consonant)}$, use STE with $threshold_2$ for correction.
- $b_p \text{ XOR } e_p == \text{(Fricative, Affricate)}$, use SBSF with $threshold_3$ for correction.
- $b_p == \text{(Unvoiced Stop Consonant)}$ and $e_p == \text{(Nasal)}$, use SBSF with $threshold_3$ for correction.
Figures: Experiments with Hindi
Syllable /t aa/
Segmentation Correction

Syllable /t aa/
Segmentation Correction

Syllable /k a r/
Segmentation Correction

Syllable /k a r/
Table: Average log probability per frame for Hindi (Female data)

Observe that vowel likelihoods have not changed significantly, while consonants, affricates change quite significantly.
### Segmentation Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Nasals</th>
<th>Fricatives</th>
<th>Affricates</th>
<th>Semi-Vowels</th>
<th>Vowels</th>
<th>Stops Consonants</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid</td>
<td>-69.02</td>
<td>-77.74</td>
<td>-79.46</td>
<td>-74.16</td>
<td>-64.49</td>
<td>-82.05</td>
<td>-70.46</td>
</tr>
<tr>
<td>HMM-FS</td>
<td>-69.80</td>
<td>-78.24</td>
<td>-80.22</td>
<td>-74.60</td>
<td>-65.07</td>
<td>-82.82</td>
<td>-71.29</td>
</tr>
</tbody>
</table>

**Table:** Average log probability per frame for Hindi (Male data)
## Segmentation Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Nasals</th>
<th>Fricatives</th>
<th>Affricates</th>
<th>Semi-Vowels</th>
<th>Vowels</th>
<th>Stops Consonants</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid</td>
<td>-69.67</td>
<td>-79.05</td>
<td>-79.98</td>
<td>-73.99</td>
<td>-67.23</td>
<td>-83.01</td>
<td>-72.99</td>
</tr>
<tr>
<td>HMM-FS</td>
<td>-70.86</td>
<td>-78.48</td>
<td>-80.40</td>
<td>-74.89</td>
<td>-68.01</td>
<td>-84.41</td>
<td>-73.93</td>
</tr>
</tbody>
</table>

**Table:** Average log probability per frame for Tamil
Boundary correction and DNN/CNNs
Proposed Approach

Combine neural networks and signal processing
Proposed Methods

- Iterative boundary correction at utterance level
DNN configuration

- 40 dimensional fbank features over 11 frames
- RBM training to initialize the DNN weights
- Stochastic gradient descent and back propagation
- mini-batch size of 256 is used
**CNN configuration**

- 40 dimensional fbank features with 3 pitch coefficients with a 11 frame context
- Convolutional window of dimension 8 and pooling window of size 3 with no-overlap.
- Feature map size of 256 and 128 used in first and second layers respectively
Architectures used

- HMM-GMM (Hybrid Segmentation) \(^{11}\)
- HMM-DNN
- HMM-DNN with boundary correction
- HMM-CNN
- HMM-CNN with boundary correction

\(^{11}\) S Aswin Shanmugam and Hema Murthy hybrid approach to segmentation of speech using group delay processing and hmm based embedded reestimation. In INTERSPEECH, 2014, pp. 1648–1651
Segmentation Correction

Figure: Example – manually marked

Syllable /sh o dh/
Segmentation Correction

Figure: Example – flat start

Syllable /sh o dh /
Segmentation Correction

Figure: Example – hybrid segmentation

Syllable /sh o dh /
Segmentation Correction

Figure: Example – CNN

Syllable /sh o dh /

Hema A Murthy (TTS Consortium, Funded by DeitY, GoI, India)
Segmentation Correction

Figure: Example – DNN

Syllable /sh o dh /

Segmentation Correction

Figure: Example – CNN + BC

Syllable /sh o dh /

Hema A Murthy (TTS Consortium, IL Speech Synthesis 03 July 2017)
Segmentation Correction

Figure: Example – DNN + BC

Syllable /sh o dh /

Syllable boundary in CLS
Flat-start
Hybrid segmentation
DNN without BC
DNN with BC
CNN without BC
CNN with BC

Waveform
Spectrogram

Syllable boundary
Time
Experiments and Results

- A subset of Indic Database is used for the experiments \(^{12}\)

**Table: Dataset used**

<table>
<thead>
<tr>
<th>Language</th>
<th>Gender</th>
<th>Duration (in hrs)</th>
<th>No. of utterances</th>
<th>No. of phones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi</td>
<td>Male</td>
<td>5.00</td>
<td>2192</td>
<td>58</td>
</tr>
<tr>
<td>Hindi</td>
<td>Female</td>
<td>5.00</td>
<td>2144</td>
<td>58</td>
</tr>
<tr>
<td>Bengali</td>
<td>Male</td>
<td>5.00</td>
<td>3093</td>
<td>52</td>
</tr>
<tr>
<td>Kannada</td>
<td>Male</td>
<td>3.43</td>
<td>1289</td>
<td>49</td>
</tr>
<tr>
<td>Kannada</td>
<td>Female</td>
<td>3.82</td>
<td>1229</td>
<td>48</td>
</tr>
<tr>
<td>Malayalam</td>
<td>Male</td>
<td>5.00</td>
<td>3063</td>
<td>52</td>
</tr>
<tr>
<td>Telugu</td>
<td>Male</td>
<td>4.24</td>
<td>2478</td>
<td>49</td>
</tr>
</tbody>
</table>

\(^{12}\)https://www.iitm.ac.in/donlab/tts/database.php
Experiments and Results I

### Table: Degradation Mean Opinion Scores (DMOS)

<table>
<thead>
<tr>
<th>Language</th>
<th>CNN</th>
<th>CNN-BC</th>
<th>DNN</th>
<th>DNN-BC</th>
<th>GMM-BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi-male</td>
<td>4.03</td>
<td>4.32</td>
<td>4.08</td>
<td>4.55</td>
<td>3.99</td>
</tr>
<tr>
<td>Hindi-female</td>
<td>3.35</td>
<td>3.70</td>
<td>3.36</td>
<td>3.51</td>
<td>3.17</td>
</tr>
<tr>
<td>Bengali-male</td>
<td>3.26</td>
<td>3.71</td>
<td>3.18</td>
<td>3.60</td>
<td>3.02</td>
</tr>
<tr>
<td>Kannada-male</td>
<td>3.64</td>
<td>3.72</td>
<td>3.42</td>
<td>3.44</td>
<td>3.40</td>
</tr>
<tr>
<td>Kannada-female</td>
<td>3.13</td>
<td>3.51</td>
<td>3.22</td>
<td>3.44</td>
<td>3.19</td>
</tr>
<tr>
<td>Malayalam-male</td>
<td>3.82</td>
<td>4.40</td>
<td>4.02</td>
<td>4.43</td>
<td>3.44</td>
</tr>
<tr>
<td>Telugu-male</td>
<td>3.50</td>
<td>4.08</td>
<td>3.67</td>
<td>3.92</td>
<td>3.48</td>
</tr>
</tbody>
</table>

### Table: Word Error Rates (%)

<table>
<thead>
<tr>
<th>Language</th>
<th>CNN</th>
<th>CNN-BC</th>
<th>DNN</th>
<th>DNN-BC</th>
<th>GMM-BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi-male</td>
<td>3.14</td>
<td>0.28</td>
<td>4.42</td>
<td>2.00</td>
<td>5.85</td>
</tr>
<tr>
<td>Hindi-female</td>
<td>7.50</td>
<td>2.50</td>
<td>6.00</td>
<td>1.00</td>
<td>8.75</td>
</tr>
<tr>
<td>Bengali-male</td>
<td>6.50</td>
<td>1.81</td>
<td>5.55</td>
<td>1.61</td>
<td>6.40</td>
</tr>
<tr>
<td>Kannada-male</td>
<td>4.33</td>
<td>2.00</td>
<td>3.33</td>
<td>2.00</td>
<td>5.66</td>
</tr>
<tr>
<td>Kannada-female</td>
<td>4.76</td>
<td>3.57</td>
<td>3.57</td>
<td>2.38</td>
<td>5.95</td>
</tr>
<tr>
<td>Malayalam-male</td>
<td>3.33</td>
<td>1.66</td>
<td>3.33</td>
<td>0.50</td>
<td>5.66</td>
</tr>
</tbody>
</table>

Hema A Murthy (TTS Consortium, funded by DeitY, GoI, India)
Sub-utterance level Approach

Boundary correction at sub-utterance level
Proposed System - Sub-utterance level

HMM-GMM flat start

Initial phone alignment

Group delay (GD) processing of STE/SBSF

Initial syllable alignment

GD corrected syllable alignment

Split wavfiles at GD corrected boundaries

HMM-GMM flat start

HMM-GMM flat start

Acoustic features

Wave files

Block I

Yes

DNN/HM/CNN-HMM training

Final boundary corrected

Phone alignment

DNN/HM/CNN-HMM training

Final boundary corrected

Phone alignment

DNN/CNN segmentation without boundary correction without iteration

DNN/CNN segmentation with boundary correction with split

Initial phone alignment

GD corrected syllable alignment

Split wavfiles at GD corrected boundaries

HMM-GMM flat start

DNN/HM/CNN-HMM training

Final boundary corrected

Phone alignment

DNN/HM/CNN-HMM training

Final boundary corrected

Phone alignment

DNN/CNN segmentation without boundary correction without iteration

DNN/CNN segmentation with boundary correction with split
Architectures used - Sub-utterance level

- HMM-DNN mono
- HMM-DNN tri
- HMM-DNN mono boundary correction at sub-utterance level
- HMM-DNN tri boundary correction at sub-utterance level
- HMM-CNN mono
- HMM-CNN tri
- HMM-CNN mono boundary correction at sub-utterance level
- HMM-CNN tri boundary correction at sub-utterance level
Figure: Sample boundary correction with DNN - Sub-utterance
### DNN/CNN Systems

<table>
<thead>
<tr>
<th></th>
<th>Flat start</th>
<th>Flat start + boundary correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>CNN + boundary correction</td>
<td></td>
</tr>
<tr>
<td>DNN</td>
<td>DNN + boundary correction</td>
<td></td>
</tr>
<tr>
<td>Bilingual DNN+BC</td>
<td>Bilingual DNN+BC</td>
<td></td>
</tr>
</tbody>
</table>

Details of this work can be found at IS 2017\(^{13}\)

---

\(^{13}\) Arun Baby et al, “Deep Learning Techniques in Tandem with Signal Processing Cues for Phonetic Segmentation for Text to Speech Synthesis in Indian Languages,” accepted IS 2017
### DNN/CNN Systems – Other languages

<table>
<thead>
<tr>
<th>Language</th>
<th>FS+BC</th>
<th>DNN+BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bengali</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bodo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gujarati</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kannada</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malayalam</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manipuri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marathi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rajasthani</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tamil</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telugu</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Issues with some of the languages: Erroneous transcription – flagging of the errors using likelihood

---

The quality is intelligible and more or less natural – naturalness needs improvement.

Synthesis is still at the sentence level. Synthesis MUST consider prosody across sentences and paragraphs.
Synthesis using syllable position alone

Figure: An example of a speech with artifacts for the text *bhaag bhii khel* which has artifact near *bhii*
Scope for Improvement

Prosody modeling based on sentence structure (NCC 2014) II

Natural sentence prosody

Figure: An example of natural speech for the text *bhaag bhii khel*
Parameters for prosody modeling I

- Intersyllable average f0 difference
- Intersyllable average duration difference
- Intersyllable average energy difference
- Syllables that are part of geminates – e.g. /bachcha/ – do not use the syllable /bach/ from this to synthesise /bach/ in /bach/ /pan/.
- Likelihood based removal of syllables.
Scope for Improvement

Parameters for prosody modeling II

Modeling duration, f0, energy

Figure: *(a) Duration difference (degrees of freedom = 14, confidence interval = 0.95), (b) Average pitch difference (degrees of freedom = 17, confidence interval = 0.95), and (c) Average energy difference (degrees of freedom = 25, confidence interval = 0.95)
Intersyllable prosody based USS

Figure: An example of speech synthesised using the above mentioned approach for the text *bhaag bhii khel*
## Example and Results II

<table>
<thead>
<tr>
<th>Language</th>
<th>DD</th>
<th>$f_0$D</th>
<th>ED</th>
<th>DD, ED, $f_0$D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi</td>
<td>3.13</td>
<td>3.63</td>
<td>3.24</td>
<td>3.29</td>
</tr>
<tr>
<td>Tamil</td>
<td>3.10</td>
<td>3.48</td>
<td>3.94</td>
<td>3.27</td>
</tr>
</tbody>
</table>

DMOS, WER
### Sample Synthesised Sentences

**Table: Synthesised Files**

<table>
<thead>
<tr>
<th>Hindi Unpruned (USS)</th>
<th>Hindi Pruned (USS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamil Unpruned (USS)</td>
<td>Tamil Pruned (USS)</td>
</tr>
<tr>
<td>Hindi HTS</td>
<td>Hindi HTS + STRAIGHT</td>
</tr>
<tr>
<td>Hindi HTS+STRAIGHT</td>
<td>Hindi HTS + Prune + STRAIGHT</td>
</tr>
</tbody>
</table>
Other efforts

- Bilingual HTS using common phone set between native language and English.
- Multilingual synthesis (SSNCE, IIITH, IITM)
- Polyglot synthesis (SSNCE, IIITH)
- Replacement of STRAIGHT in HTS (IISc)
Applications

- Highlighting of text on the web
- Online examination for visually challenged
- Small footprint TTS ported to i) Android based systems ii) communication devices like HOPE/KAVI – Chetana (NGO’s) products for cerebral palsy patients.
  - Ported to Samsung S5360 Galaxy Y, Samsung Tablet, Adam (notion Ink), Akash Tablet
  - Integration with different Android based products developed at IIT Mandi
  - Integration with OCR (CDAC, Tvm)
  - Integration with ASR for crop information (CDAC, Mumbai)
Awards won and other efforts

- Training Visually Challenged persons on Word Processor, Spreadsheet, e-mail, Internet: 5 years, 180 persons.
- Manthan Award: 2012 (Top 74 finalists), 2012.
- GE Innovation and Research Expo Award: 2013.
- Launch of SMS Reader in 5 Indian languages, February 2014.
- Launch of TTS systems in 9 Indian languages on good-governance day by Hon’ble Minister of IT, Dec 2015.
- Freely available on the web: www.iitm.ac.in/donlab/tts/

Most importantly: Please check our websites

http://www.iitm.ac.in/donlab/hts/ – statistical parametric synthesisers
http://www.iitm.ac.in/donlab/festival/ – unit selection synthesisers

Give your Feedback
Thank you for your Attention

Acknowledgements: Prof. B Yegnanarayana, Prof. K V Madhu Murthy, Prof.C Chandra Sekhar and ALL my students who TAUGHT/TEACH me.
Mark your calendars – see you at INTERSPEECH 2018
Sept 2-6, 2018, Hyderabad HICC, India