2. DATA

Data Source: Indic TTS data, TTS Consortium, Funded by Govt of India

Language: Hindi
No. of Speakers: 1 Female
No. of Sentences: 2318
Total Hours: 5.2
Avg. Audio Length: 8 sec
Avg. Sentence Len: 97 char
Train/Test (%) : 80/15/5
Character set: 71

3. ARCHITECTURE: FULLY CONVOLUTIONAL TTS WITH GUIDED ATTENTION & SINUSODIAL POSITION ENCODING

JOURNEY OF A CHARACTER AND AUDIO SIGNAL
- Sentence -> Character -> Character Embedding -> Encoded Text
- Audio -> Mel-Spectrogram -> Encoded Audio -> Align with Encoded Text -> Predicted Mel -> Train Text2Mel
- Audio -> Spectrogram (ground-truth Spec) vs. Mel -> Predicted Spec, to train SSBN

ARCHITECTURE FEATURES
- Stacked Dilated Conv1D -> Gain context information, Faster Training
- Highway Activations -> Manage vanishing gradients
- Guided Attention -> Penalize non-diagonal attention matrix
- Position Encoding -> Reduce attention errors

4. MODEL VARIATIONS

<table>
<thead>
<tr>
<th>Model</th>
<th>Val performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1 (DC-TTS, dropout 40%)</td>
<td>Attention does not converge</td>
</tr>
<tr>
<td>M2 (M1 + Position Encoding + Learned Embedding)</td>
<td>NWR = 73.1%</td>
</tr>
<tr>
<td>M3 (M1 + Position Encoding + Pretrained Embedding)</td>
<td>Subjective poor than M2</td>
</tr>
</tbody>
</table>

5. MODEL PERFORMANCE METRIC

- Synthesized Audio -> Google Speech-To-Text -> Compare Synthesized Text with Actual Text -> Synthesized Word Error Rate (WER)
- Normalized by Google STT accuracy for groundtruth Audio

Normalized Word Recognition Rate = 1 - Synthesized Audio WER
L = Groundtruth Audio WER

6. RESULTS FOR FINAL MODEL

<table>
<thead>
<tr>
<th>Data</th>
<th>Norm. WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>73.1%</td>
</tr>
<tr>
<td>Test</td>
<td>77.5%</td>
</tr>
<tr>
<td>Short Sentence</td>
<td>65%</td>
</tr>
<tr>
<td>Medium Length Sentence</td>
<td>72%-74%</td>
</tr>
<tr>
<td>Long Sentences</td>
<td>~78%</td>
</tr>
</tbody>
</table>
- Good accuracy, naturalness and alignment for medium to long sentences

7. ATTENTION AND LOSS

8. OTHER OBSERVATIONS

9. SUMMARY, FUTURE WORK

- Hindi TTS with no hand-engineered features and fast training that works reasonably well even with small training dataset
- HP Tying, Mixed Character-to-Phoneme model, training longer can improve further
- Extend model to multi-speaker, multi-lingual and code-mixed data

10. SELECTED REFERENCES


Poster Video shared at: https://youtu.be/WwZABH5ZJ9Q