1. MOTIVATION

4th — Most Spoken Language in World

129mn — # English Speakers in India, but

528mn, 270 — # Hindi Speakers in India, # Mother Tongues

272mn, 62mn — # Illiterates, # Visually Impaired

Indic Languages need due 'Attention'

Listen, Speak in Hindi!

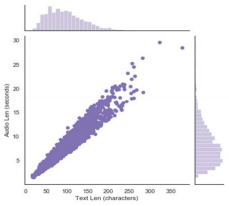
Text-to-Speech Synthesis for Hindi

Dinesh Chaudhary, chdinesh@stanford.edu, SUID: 06349844

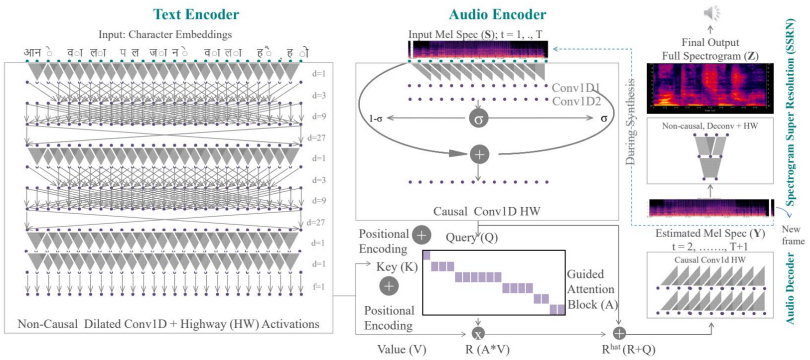
CS-230, Deep Learning, Winter 2019, Final Project

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/Val/Test (%)	80/15/5
Character set	71



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



SSRN Loss
L1 Loss
+
Binary Divergence Loss b/w **Z** & ground truth Spec.

Text2Mel Loss
L1 Loss
+
Binary Divergence Loss b/w **Y** & **S**

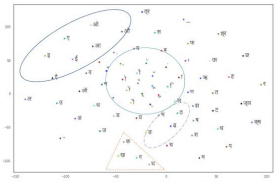
JOURNEY OF A CHARACTER AND AUDIO SIGNAL

- **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 SSRN

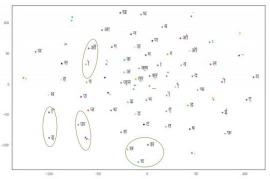
ARCHITECURE FEATURES

- 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

8. OTHER OBSERVATIONS



Pre-trained Character Embedding



Learned Character Embedding

- Both pretrained and learned character embedding show proximity to other characters with similar sound or origin source, though degree of overlap is low.

4. MODEL VARIATIONS

Model	Val performance
M1 (DC-TTS, dropout 40%)	Attention does not converge
M2 (M1 + Position Encoding + Learned Embedding)	NWRR = 73.1%
M3 (M1 + Position Encoding + Pretrained Embedding)	Subjective assessment, poor than M2

5. MODEL PERFORMANCE METRIC

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

Normalized Word Recognition Rate =

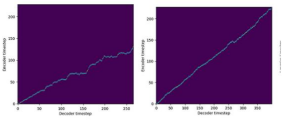
$$\frac{1 - Synthesized\ Audio\ WER}{1 - Groundtruth\ Audio\ WER}$$

6. RESULTS FOR FINAL MODEL

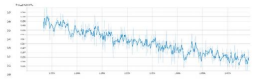
Data	Norm. WRR
Validation	73.1%
Test	77.5%
• Short Sentence	60%
• Medium Length Sentence	72%-74%
• Long Sentences	~78%

- Good accuracy, naturalness and alignment for medium to long sentences

7. ATTENTION AND LOSS



10K iteration — 30K iteration



Loss per Epoch

9. SUMMARY, FUTURE WORK

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

10. SELECTED REFERENCES

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin (2017) Attention is all you need
- Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., and Miller, J. (2018) Deep Voice 3: 2000-speaker neural text-to-speech.
- Tachibana, H., Uenoyama, K., and Aihara, S. (2018) Efficiently trainable TTS system based on deep

Poster Video shared at: https://youtu.be/WwZA0H5Z3O4