Leveraging Effective Language and Speaker Conditioning in Indic TTS for LIMMITS 2024 Challenge

Yejin Jeon*, Youngjae Kim*, and Gary Geunbae Lee

Task Definition

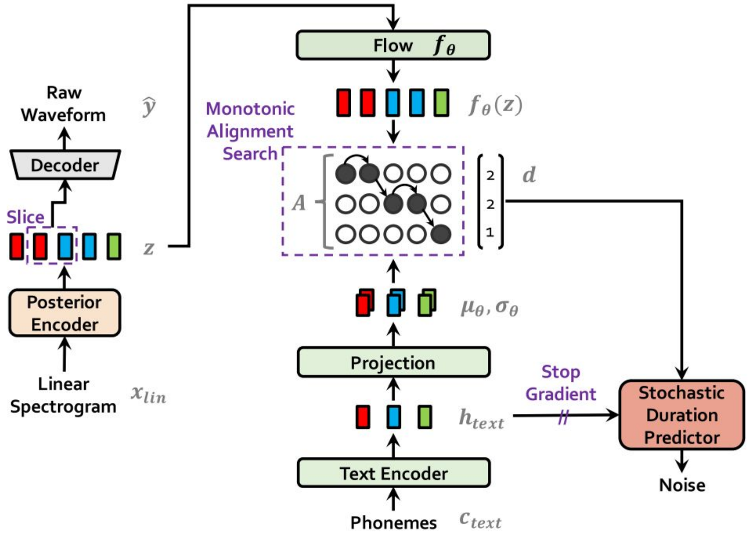
Track 1 - Few shot TTS+VC with challenge dataset

Using a pretrained multi-lingual, multi-speaker TTS built on the challenge dataset, perform few shot voice cloning by fine-tuning new speakers.

Emi chesthunnaru?
(TELUGU)

तुम्ही मला मदत करू शकता का?
(MARATHI)

Preliminaries

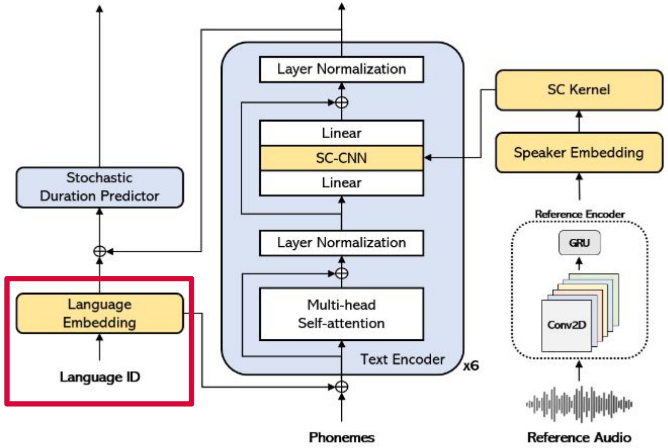


[Baseline]

- Utilizes an adversarial autoencoder to generate similar distributions between the phoneme representation and reference audio.

- End-to-end (E2E) one-stage paradigm
  - For easier / efficient training

Methodology



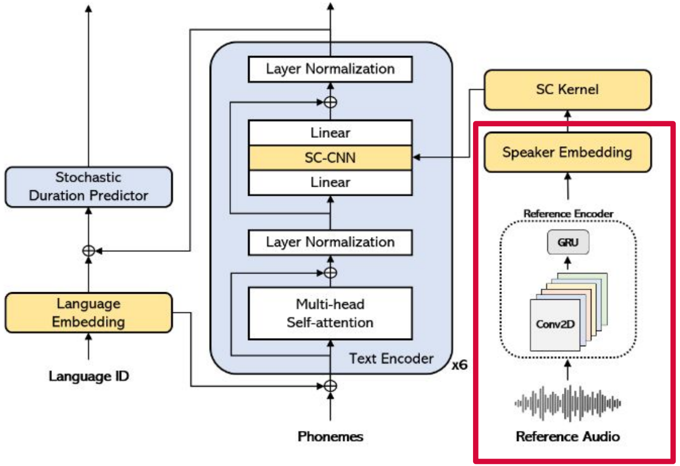
[Multi-lingual Settings]

- Language Embedding
  - Language ID Alignment, and conversion into 256 dimensional vector

- Integration of Language Information
  - Concatenation with phoneme embedding at the beginning of the text encoder.
  - Concatenation with text encoder outputs, which is used as inputs for stochastic duration predictor.

- Language embeddings go through additional conv1d layer for integration with hidden states.

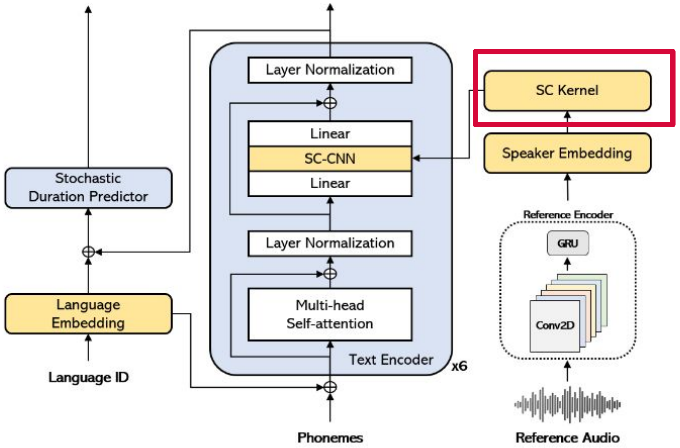
Methodology



[Multi-speaker Settings]

- Mel-spectrograms that are converted from reference audio are passed to a reference encoder made up of six 2-D convolution layers of filters [32, 32, 64, 64, 128, 128], and a GRU layer.
    - Results in initial speaker embedding *s*

Training Scheme

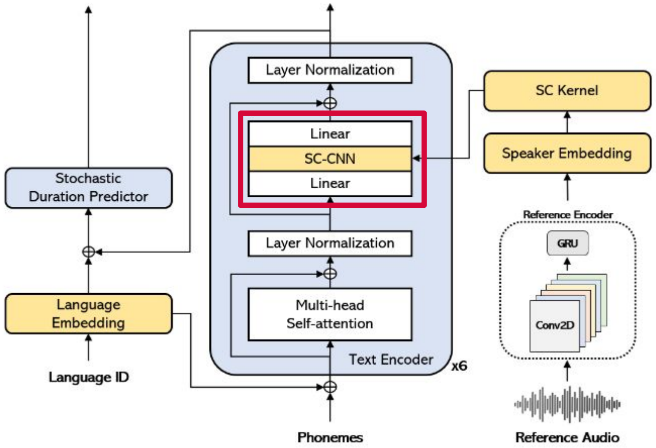


[Multi-speaker Settings]

- A single linear layer is used to extract the weights and biases ("kernel variables") from speaker embedding *s.*

$$\{D_{dir}, D_{gain}, D_{bias}\} = Linear(s)$$

Training Scheme

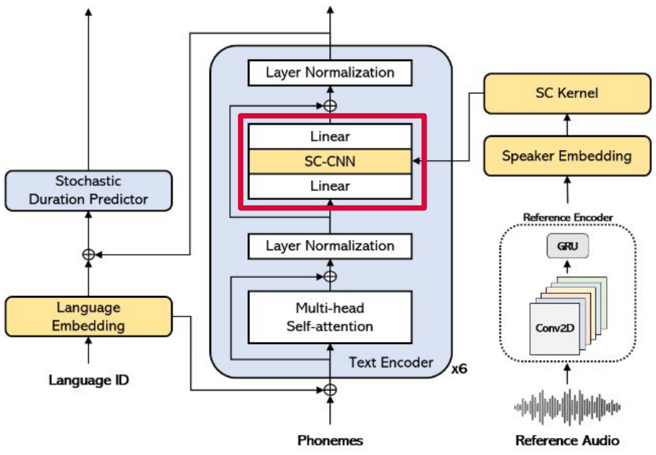


[Multi-speaker Settings]

- One conv1d layer is used to fuse the speaker kernel variables with the phonemic representations.

$$Fusion = (P_{gain}\frac{P_{dir}}{||P_{dir}||}) * ((D_{gain}\frac{D_{dir}}{||D_{dir}||}) * x + D_{bias}) + P_{bias} \quad (2)$$

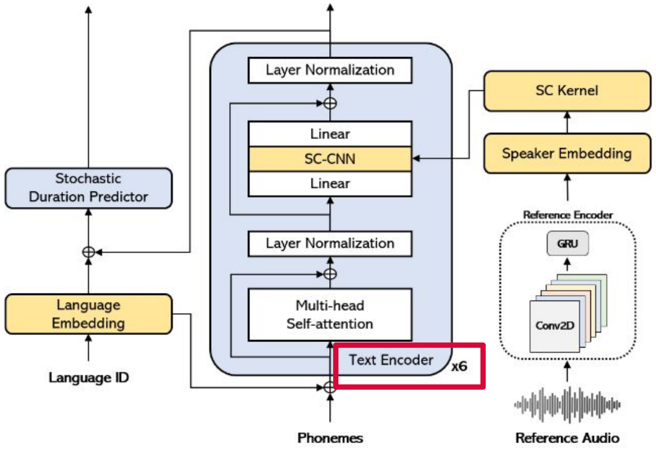
Training Scheme



[Multi-speaker Settings]

- Original Transformer CNN layers are substituted with linear layers, with the speaker-related convolution layer placed in between.

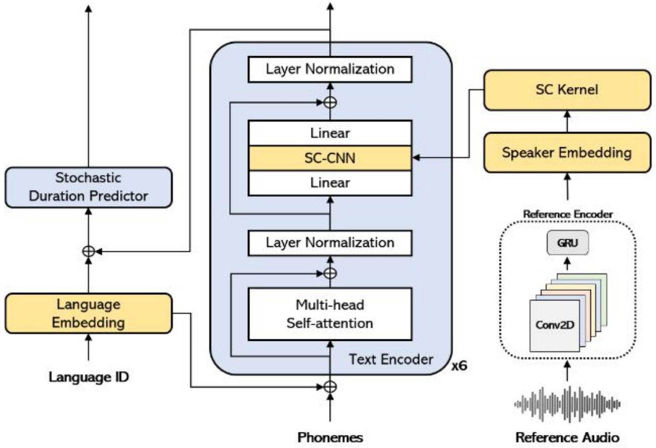
Training Scheme



[Training Settings]

- Speaker information is only integrated starting from the third iteration of the text encoder (Reference [2]).
    - The outputs from the first two iterations are directly passed to the duration predictor in order to generate speaker-independent durations.

Training Scheme



[Training Settings]

$$L_{vae} = L_{recon} + L_{kl} + L_{dur} + L_{adv}(G) + L_{fm}(G)$$

Equation from VITS paper

Experimental Settings

[Original LIMMITS Dataset]

- 14 speakers of equal gender distribution across 7 different languages
- 560 hour corpus

Experimental Settings

[Original LIMMITS Dataset]

- 14 speakers of equal gender distribution across 7 different languages
- 560 hour corpus

[Partial LIMMITS Dataset]

- 1 hour per speaker
- A total of 14 hours with an average of 16.17 words per audio sample
- 22050 Hz

Experimental Settings

[Common Settings]

- 75 million parameters
- 4 NVIDIA A100 GPUs
- 64 batch size

[Pre-Training]

- 410k steps over a span of 3 days

[Fine-Tuning]

- 90k steps over a span of 18 hours

Model Optimization

Training Epochs		CER
50000 iteration	English	8.6%
	Hindi	14.93%
90000 iteration	English	8.5%
	Hindi	15.09%
115000 iteration	English	9.77%
	Hindi	15.02%

- Further training does not necessitate in better performance.
- Clear pronunciation errors for English when training models for a longer period of time.

Official Results

Table 1. Results for naturalness and speaker similarity.

	Average	σ
Naturalness (MOS)	3.74	1.02
Similarity (Score)	3.85	1.34

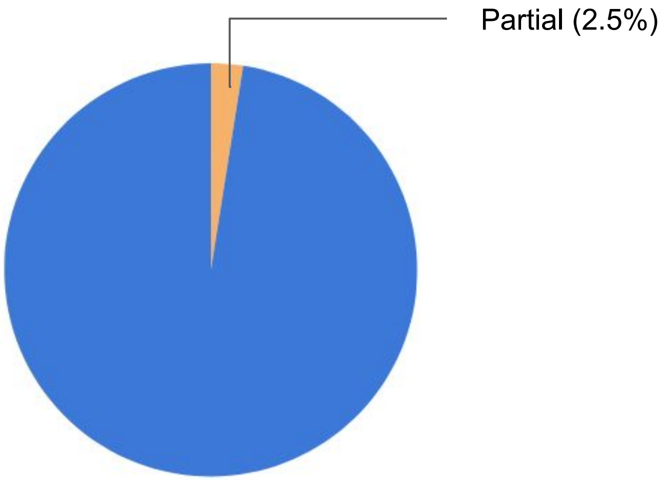
- The submissions will be evaluated on naturalness and speaker similarity scores, for mono lingual and cross lingual synthesis.
- Each submission will be evaluated by multiple evaluators, native to the target language.

- *From the LIMMITS Website*

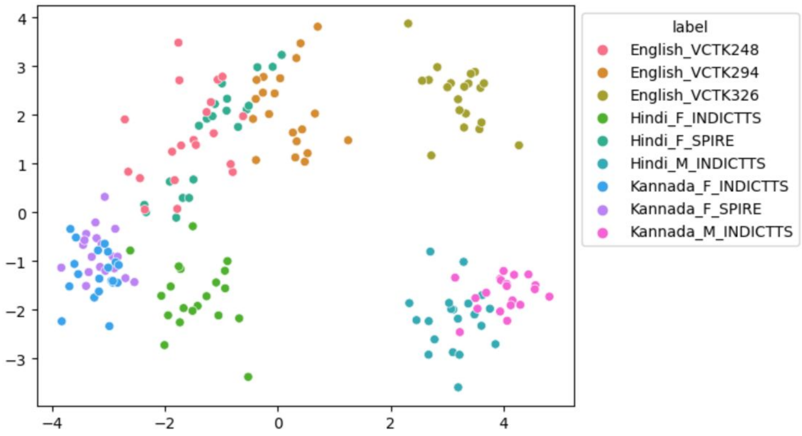
Research Questions

Analysis - Speaker

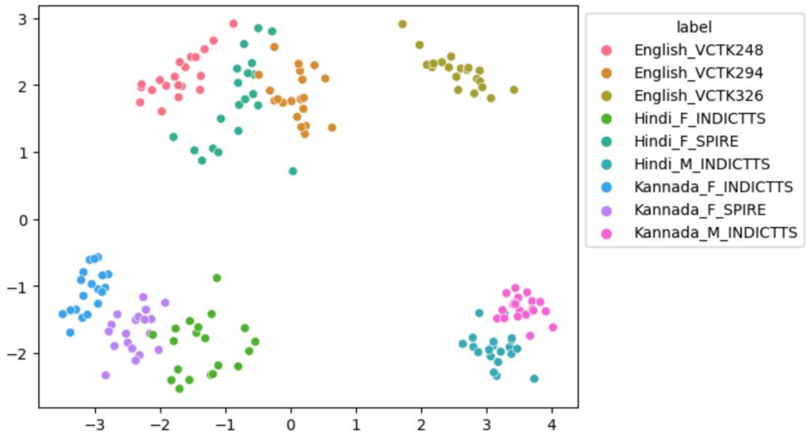
Q) Is there a difference between using the partial and full LIMMITS dataset?



Analysis - Speaker



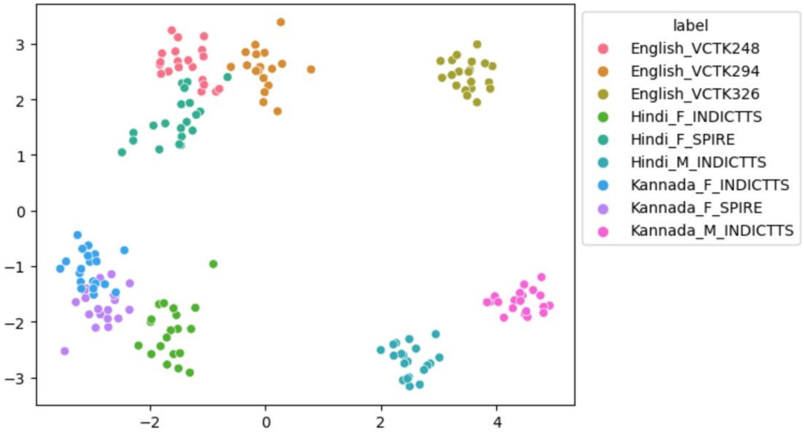
Pretrained 1 hour / speaker dataset



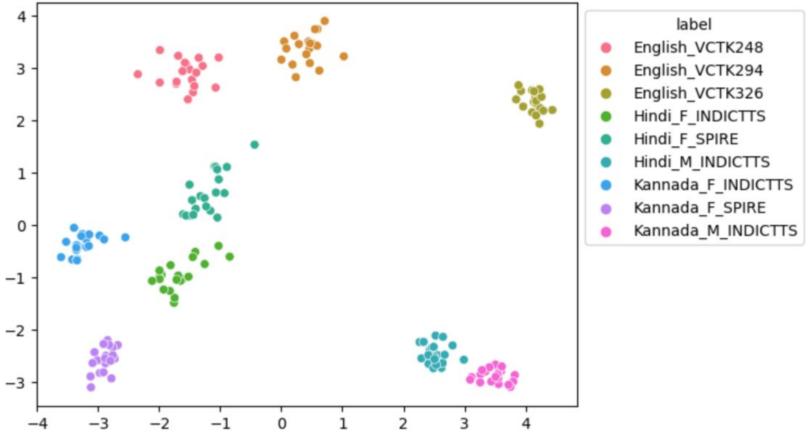
Pretrained Full dataset

- Model pre-trained using 14-hour corpus results in speaker embeddings that are relatively more scattered compared to the same model that was pre-trained on the full 560-hour corpus.

Analysis - Speaker



Fine Tuning
from 1 hour / speaker dataset

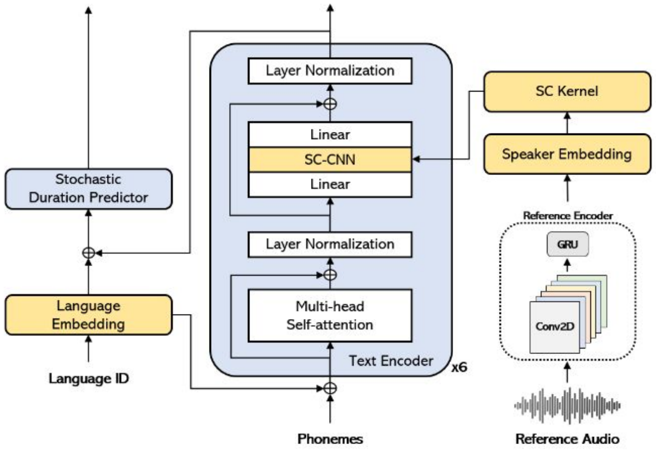


Fine Tuning
from Full dataset

- Fine tuning models trained on the partial and full dataset results in similar speaker embeddings.
- Not much of a difference between partial and full dataset utilization in terms of speaker distinguishment.

Research Question #2

Analysis - Speaker



[Training Settings]

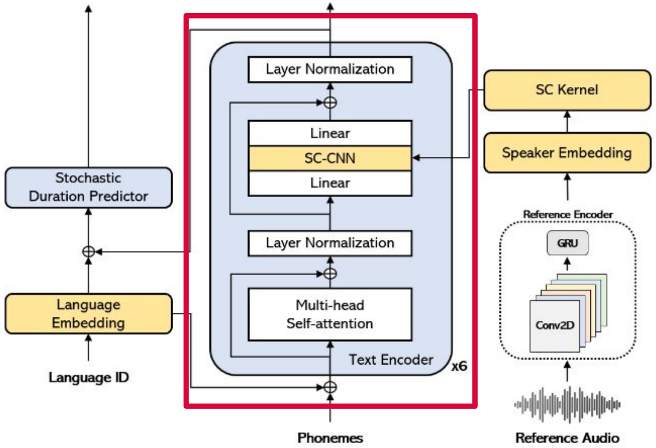
- Speaker information is only integrated starting from the third iteration of the text encoder (Reference [2]).

Analysis - Speaker

Q) Will additional speaker information integration improve performance?

versus

Analysis



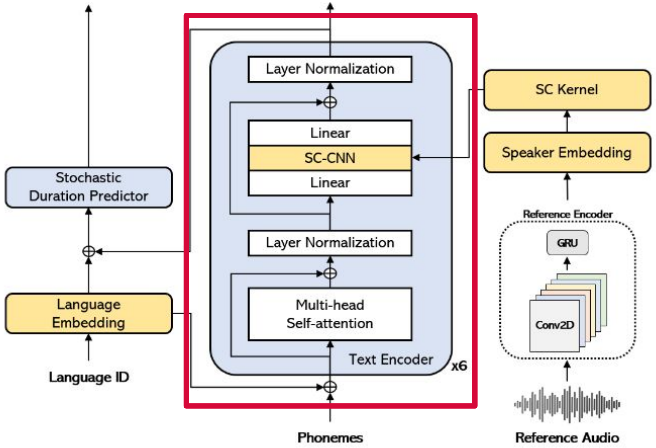
[Additional Speaker Fusion]

- Integrated SC Kernels into the last 4 / 6 / 8 iterations when there were 6 / 8 / 10 text encoder blocks, respectively.

- Conducted mono- and cross-lingual MOS for audios synthesized in the target language of English.
    - No native speakers for other Indic languages

- No significant differences in terms of speaker similarity.

Analysis



[Speaker Fusion]

- Pre-trained Whisper2
  - Conducted only for English and Hindi

- 10 iterations shows the best CER scores for Hindi.

Layers		CER
6 Iterations	English	8.45% ± 0.61
	Hindi	15% ± 0.37
8 Iterations	English	9.63% ± 0.91
	Hindi	15.51% ± 0.76
10 Iterations	English	9.27% ± 0.99
	Hindi	**14.88% ± 0.52**

Analysis

Layers		CER
6 Iterations	**English**	**8.45% ± 0.61**
	Hindi	**15% ± 0.37**
8 Iterations	English	9.63% ± 0.91
	Hindi	15.51% ± 0.76
10 Iterations	English	9.27% ± 0.99
	Hindi	14.88% ± 0.52

[Speaker Fusion]

- Only using 6 iterations for the text encoder demonstrates better and stable performance for both English and Hindi.

- No significant results to back reasons for utilizing additional speaker information fusion.
  - Use settings leading to overall lower CER and less model parameters.

Conclusion

- Simple, but effective language and speaker information integration methodology.
- Just using a 14-hour partial dataset results in natural and high speaker fidelity for both mono- and cross-lingual settings.

References

[1] Jaehyeon Kim, Jungil Kong, and Juhee Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in International Conference on Machine Learning.

[2] Hyungchan yoon, Changhwan Kim, Seyun Um, Hyun-Wook Yoon, and Hong-Goo Kang, “SC-CNN: Effective Speaker Conditioning Method for Zero-Shot Mult-Speaker Text-to-Speech Systems,” in IEEE Signal Processing Letters, 2023.