INTERNATIONAL AUDIO LABORATORIES ERLANGEN
A joint institution of Fraunhofer IIS and Universität Erlangen-Nürnberg

Low-Resource Text-to-Speech Synthesis Using Noise-Augmented Training of ForwardTacotron

Kishor Kayyar Lakshminarayana, Frank Zalkow, Christian Dittmar, Nicola Pia, Emanuël Habets

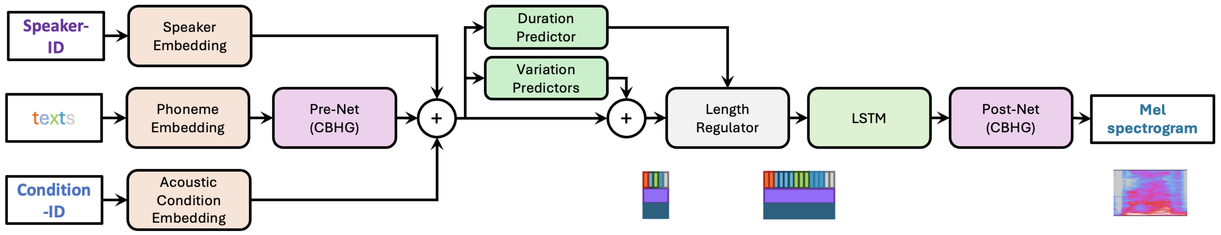
kishor.kayyar.lakshminarayana@iis.fraunhofer.de

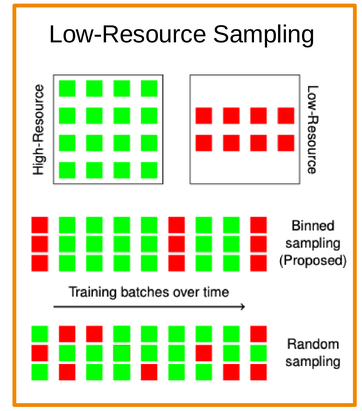
Abstract

In recent years, several text-to-speech systems have been proposed to synthesize natural speech in zero-shot, few-shot, and low-resource scenarios. However, these methods typically require training with data from many different speakers. The speech quality across the speaker set typically is diverse and imposes an upper limit on the quality achievable for the low-resource speaker. In the current work, we achieve high-quality speech synthesis using as little as five minutes of speech from the desired speaker by augmenting the low-resource speaker data with noise and employing multiple sampling techniques during training. Our method requires only four high-quality, high-resource speakers, which are easy to obtain and use in practice. Our low-complexity method achieves improved speaker similarity compared to the state-of-the-art zero-shot method HierSpeech++ and the recent low-resource method AdapterMix while maintaining comparable naturalness. Our proposed approach can also reduce the data requirements for speech synthesis for new speakers and languages.

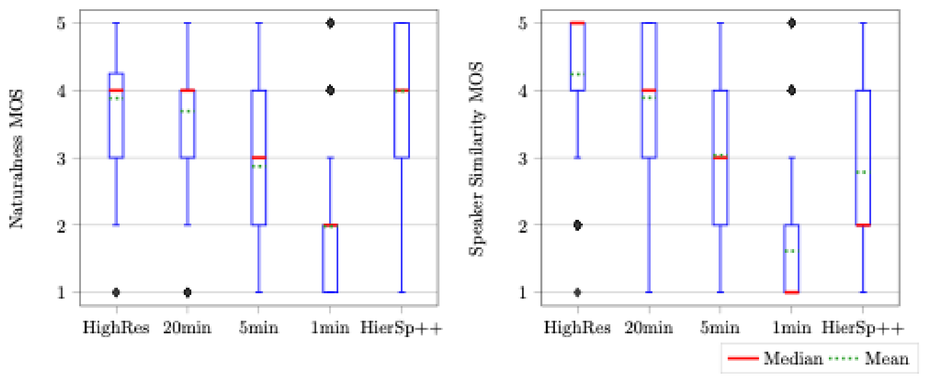
Proposed Method

- ForwardTacotron [1] with prosodic variation predictors is our basis [2].
- Four high-resource (HR) and one low-resource (LR) speakers.
- The LR training data is augmented using white Gaussian noise, similar to Kayyar, et al. [3].
- Multiple augmented versions for each LR training sample.
- We use acoustic condition embeddings (via condition-IDs).
- Training is conducted using weighted and binned sampling strategies.
- HR samples grouped into a HR bin and LR samples grouped to a LR bin.
- Training batches selected from the two bins at the rate proportional to their sizes – binned sampling.





Results



Model or LR subset used	MCD (↓)	cos-sim (↑)
High-Resource	44.6 ± 0.8	0.65 ± 0.009
HierSpeech++ [4]	54.9 ± 0.6	0.30 ± 0.008
Proposed (20 min)	50.6 ± 0.8	0.56 ± 0.009
Proposed (5 min)	53.7 ± 0.8	0.47 ± 0.012
Proposed (1 min)	59.7 ± 1.1	0.34 ± 0.028

https://s.fhg.de/ftlrtts

Conclusions

- Proposed a light-weight approach for low resource speech synthesis using noise augmentation and modified sampling strategies.
- We get improved speaker similarity at similar naturalness to state-of-the-art zero-shot method.
- Our approach trained with as little as 20 minutes of data is inferior by only 0.2 MOS regarding both naturalness and speaker similarity to a model trained with 5.5 hours of data.

References

[1] C. Schäfer, O. McCarthy, and contributors, "ForwardTacotron," https://github.com/as-ideas/ForwardTacotron, 2020.

[2] F. Zalkow, P. Sani, M. Fast, J. Bauer, M. Joshaghani, K. Kayyar, E. A. P. Habets, and C. Dittmar, "The AudioLabs system for the Blizzard Challenge 2023," in Proceedings of the Blizzard Challenge Workshop, Grenoble, France, 2023, pp. 63–68.

[3] K. Kayyar, C. Dittmar, N. Pia, and E. Habets, "Low-resource text-to-speech using specific data and noise augmentation," in Proc. IEEE-SPS European Signal Processing Conf., 2023, pp. 61–65

[4] S.-H. Lee, H.-Y. Choi, S.-B. Kim, and S.-W. Lee, "Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis," arXiv preprint arXiv:2311.12454, 2023.