HOW SHOULD *I* PRONOUNCE *YOUR* NAME?

Marius Fleischer, Aditya Gulati, Avani Tanna

mariusfleischer, adityagulati, avani@ucsb.edu

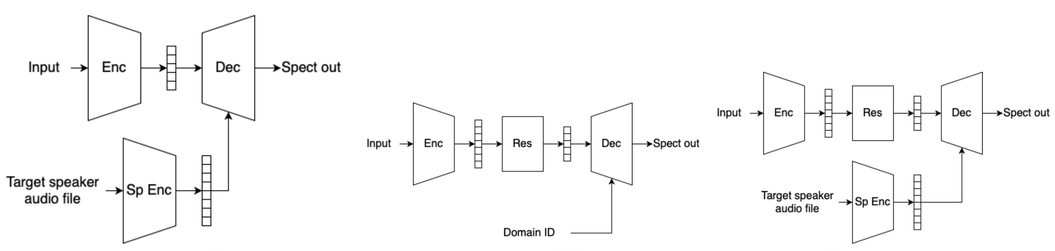
Abstract

Speech style transfer is defined as a task where the prosody, accent, pitch and voice of an individual is overlaid on another person's audio clip. If you have an international/foreign/unconventional name, you know how difficult it can be to have people pronounce it correctly! In order to solve this challenging problem and improve **pronunciations of names**, we explore speech style transfer models. The S.O.T.A. has been approached using **autoencoders** in AutoVC [2] (for zero-shot conversions) and AutoPST [1] (for prosody disentanglement). We combine the techniques in the two papers to get **zero-shot conversion** and **prosody disentaglement**.

Contributions

Architecture

Equip AutoPST with a speaker encoder that extracts **true speaker representation**.

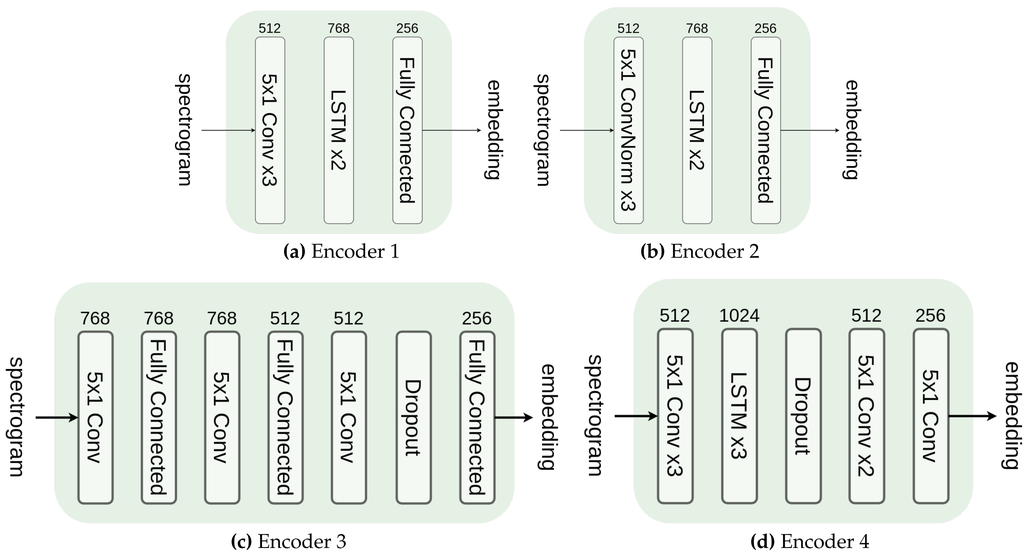


(a) AutoVC **(b)** AutoPST **(c)** Our Architecture

Figure 1: Architecture

Speaker Encoders

Design 4 custom speaker encoders of **different architectures** and **complexities**.

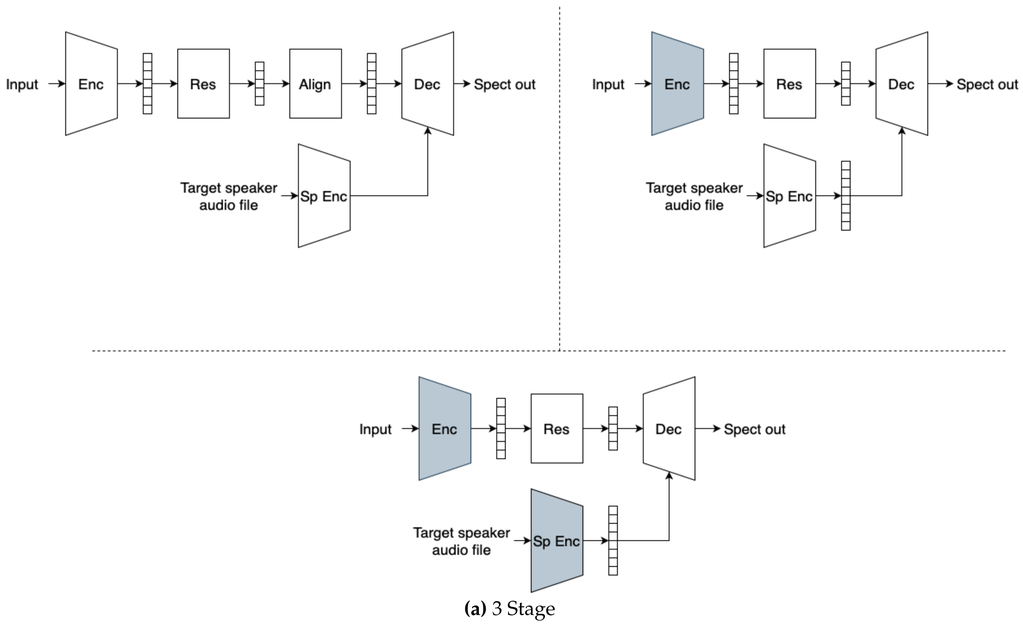


(a) Encoder 1 **(b)** Encoder 2 **(c)** Encoder 3 **(d)** Encoder 4

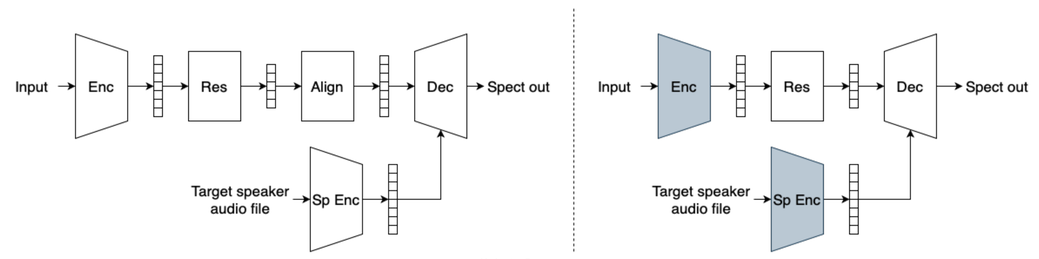
Figure 2: Speaker Encoders

Training

Design **2 stage** and **3 stage training paradigms** to train our architecture.



(a) 3 Stage



(b) 2 Stage

Figure 3: Training

Experiments

We run the 2 stage training paradigm for 1K, 2K, and 5K iterations on our 4 custom encoders. Ideally, the complete training takes 1M iterations.

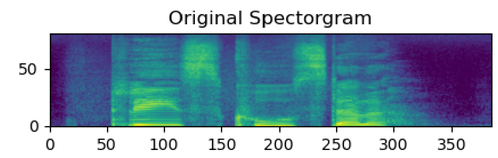
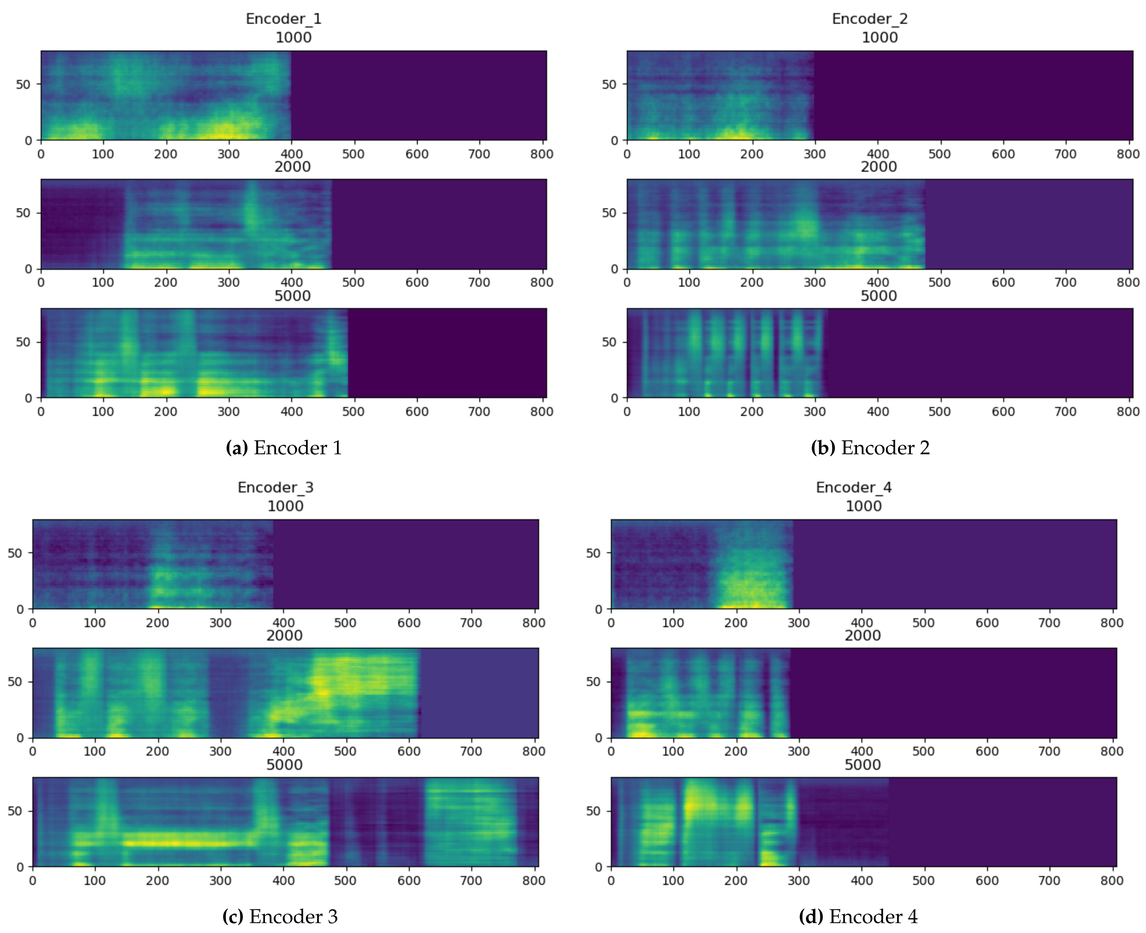


Figure 4: Original Spectrogram



(a) Encoder 1 **(b)** Encoder 2 **(c)** Encoder 3 **(d)** Encoder 4

Figure 5: Speaker Encoders results

Oberservations and Takeaways

- Spectrogram evolves similarly as AutoPST output, hence showing **positive evidence towards our approach for zero-shot** working.
- Prosody disentanglement (especially in AutoPST) works well for **accent transfer**.
- Encoder 1 performs best for less rounds of training but **Encoder 3** potentially performs best after finishing training.
- **Check out our live demo - let's pronounce your name!**

References

[1] K. Qian, Y. Zhang, S. Chang, J. Xiong, C. Gan, D. Cox, and M. Hasegawa-Johnson. Global rhythm style transfer without text transcriptions, 2021.

[2] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson. AutoVC: Zero-shot voice style transfer with only autoencoder loss. volume 97 of *Proceedings of Machine Learning Research*, pages 5210–5219, Long Beach, California, USA, 09–15 Jun 2019. PMLR.