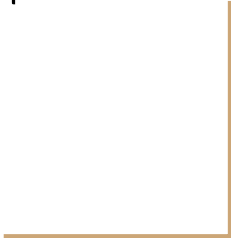




Voice Cloning

Computer Vision Reading Group (IITK)

March 22, 2022



Neural Voice Cloning with a Few Samples

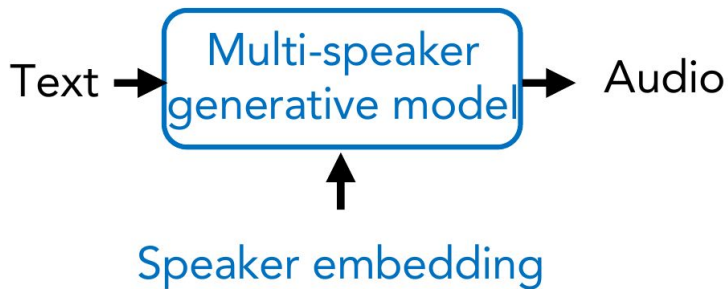
Arik, Sercan O., Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou
(NeurIPS 2018)

<https://arxiv.org/abs/1802.06006v3>

Introduction

- Multi-speaker Speech synthesis
 - Synthesizing speech conditioned on **text** and **speaker identity**
 - **Text:** linguistic information
 - **Speaker identity:** characteristics such as pitch, speech rate and accent
 - Can only generate speech for seen speakers
- Voice cloning
 - Learning the voice of an unseen speaker from a few speech samples
 - Challenges:
 - Learning speaker characteristics from limited amount of data
 - Generalize to unseen texts

Multi-speaker Speech Model



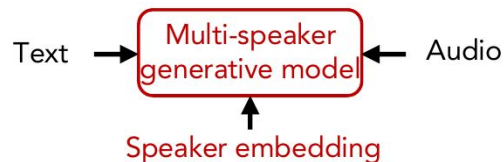
- Multi-speaker generative modeling

$$\min_{W, \mathbf{e}} \mathbb{E}_{s_i \sim \mathcal{S}, (\mathbf{t}_{i,j}, \mathbf{a}_{i,j}) \sim \mathcal{T}_{s_i}} \{L(f(\mathbf{t}_{i,j}, s_i; W, \mathbf{e}_{s_i}), \mathbf{a}_{i,j})\}$$

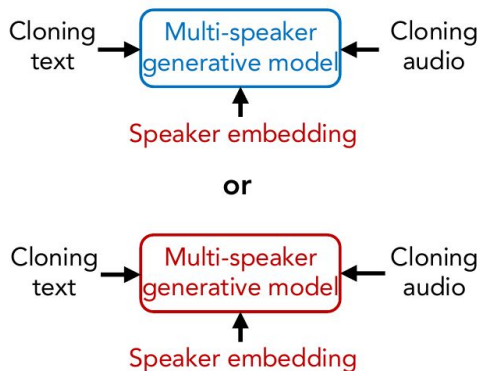
Voice Cloning Pipeline

- Voice Cloning:
 - **Training:** Train a multi-speaker generative model on seen speakers
 - **Cloning:** Extract speaker characteristics of an unseen speaker from a set of cloning audios
 - **Audio Generation:** Generate an audio given any text for that speaker
- Two approaches
 - Speaker adaptation
 - Speaker Encoding

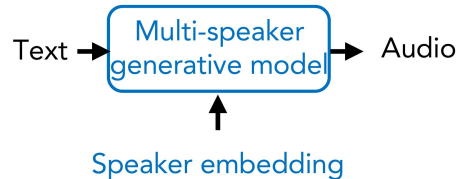
Speaker Adaptation



Training



Cloning



Audio Generation

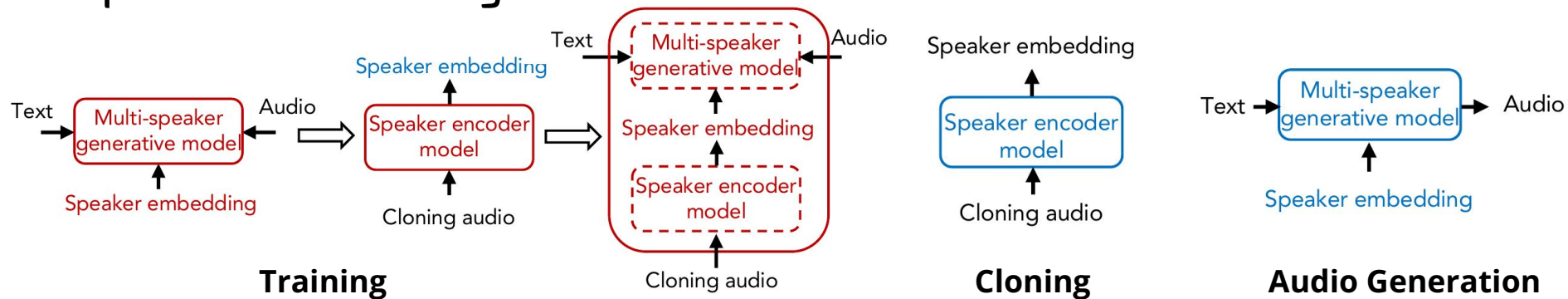
- Fine-tune a trained multi-speaker model for unseen speakers using a few audio-text pairs
- Embedding-only adaptation

$$\min_{\mathbf{e}_{s_k}} \mathbb{E}_{(\mathbf{t}_{k,j}, \mathbf{a}_{k,j}) \sim \mathcal{T}_{s_k}} \left\{ L \left(f(\mathbf{t}_{k,j}, s_k; \widehat{W}, \mathbf{e}_{s_k}), \mathbf{a}_{k,j} \right) \right\}$$

- Whole model adaptation

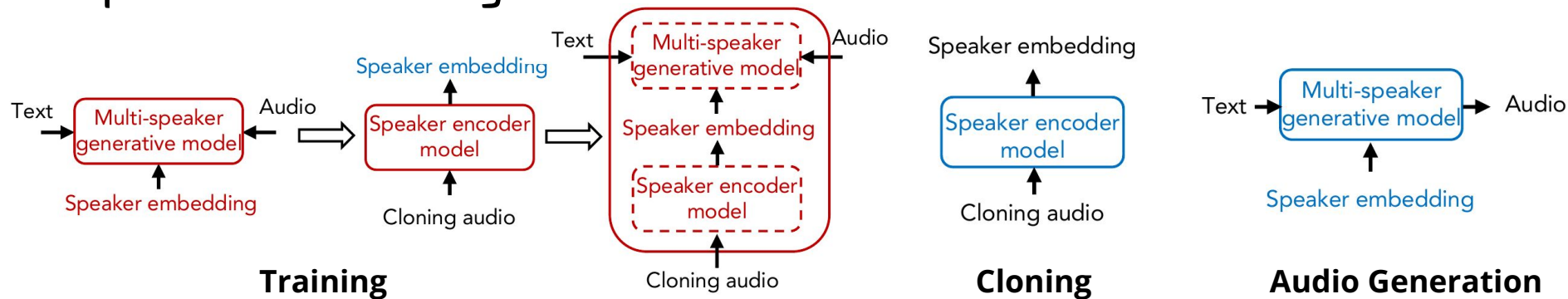
$$\min_{W, \mathbf{e}_{s_k}} \mathbb{E}_{(\mathbf{t}_{k,j}, \mathbf{a}_{k,j}) \sim \mathcal{T}_{s_k}} \left\{ L \left(f(\mathbf{t}_{k,j}, s_k; W, \mathbf{e}_{s_k}), \mathbf{a}_{k,j} \right) \right\}$$

Speaker Encoding



- Uses a speaker encoder
 - Estimates speaker embeddings from a set of cloning audio samples

Speaker Encoding



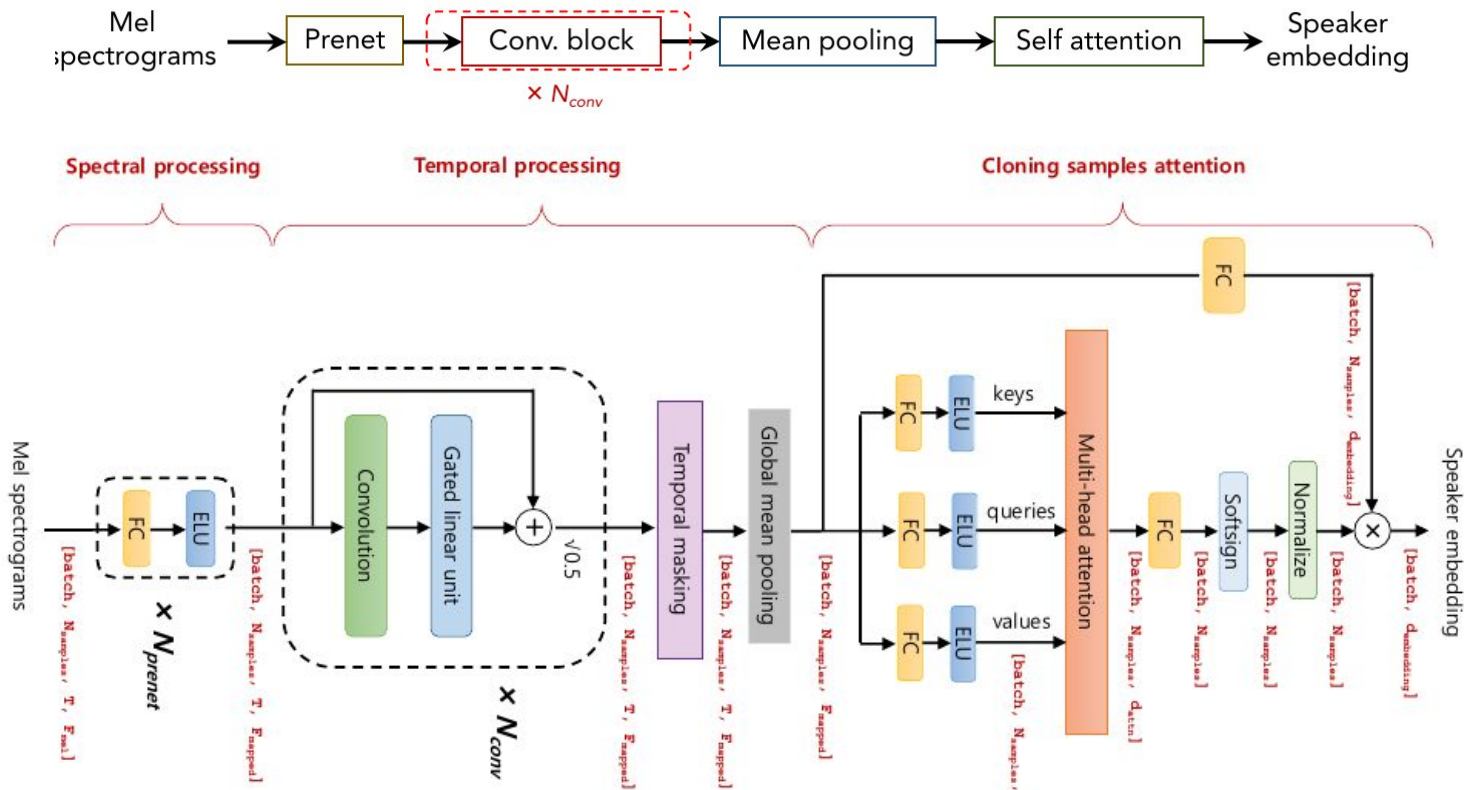
- **Training Step 1:** Train multi-speaker generative model and extract the speaker embeddings
- **Training Step 2:** Train speaker encoder with an L1 loss to predict the trained embeddings from cloning audios:

$$\min_{\Theta} \mathbb{E}_{s_i \sim \mathcal{S}} \{ |g(\mathcal{A}_{s_i}; \Theta) - \hat{\mathbf{e}}_{s_i}| \}$$

- **Training Step 3:**
Joint fine-tuning:

$$\min_{W, \Theta} \mathbb{E}_{\substack{s_i \sim \mathcal{S}, \\ (\mathbf{t}_{i,j}, \mathbf{a}_{i,j}) \sim \mathcal{T}_{s_i}}} \{ L(f(\mathbf{t}_{i,j}, s_i; W, g(\mathcal{A}_{s_i}; \Theta)), \mathbf{a}_{i,j}) \}$$

Speaker encoder Architecture



Experiments

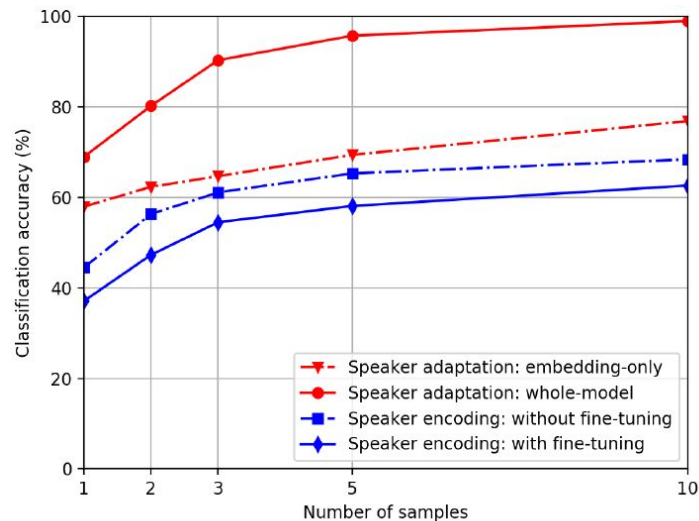
- Datasets
 - Multi-speaker generative model and speaker encoders are trained using LibriSpeech dataset (2484 speakers, 820 hours)
 - Voice cloning is performed on VCTK dataset (108 speakers, 44 hours)
- Model details
 - Deep Voice 3 with Griffin Lim Decoder
- Evaluation
 - Speaker Classification (Accuracy %)
 - A speaker classifier is trained with the set of speakers used for cloning
 - Speaker Verification (EER %)
 - Binary classification to identify whether a test audio and an enrolled audio are from the same speaker
 - Can be used for unseen speakers

Time and Memory Footprint

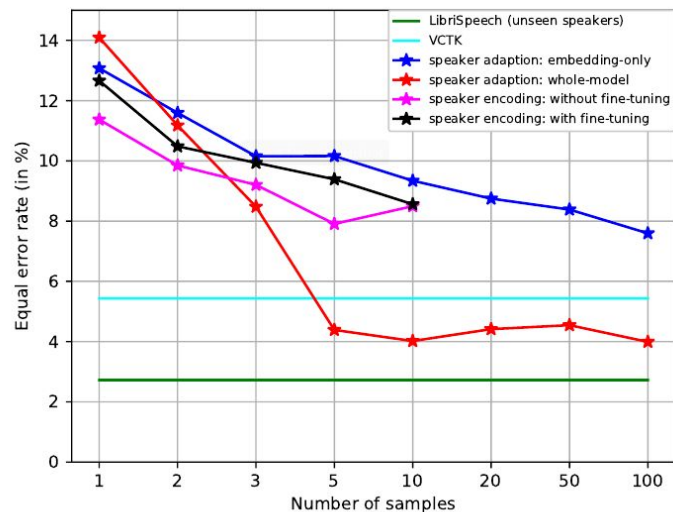
	Speaker adaptation		Speaker encoding	
Approaches	Embedding-only	Whole-model	Without fine-tuning	With fine-tuning
Data	Text and audio		Audio	
Cloning time	~ 8 hours	$\sim 0.5 - 5$ mins	$\sim 1.5 - 3.5$ secs	$\sim 1.5 - 3.5$ secs
Inference time	$\sim 0.4 - 0.6$ secs			
Parameters per speaker	128	~ 25 million	512	512

Table 1: Comparison of speaker adaptation and speaker encoding approaches.

Speaker classification accuracy and verification EER



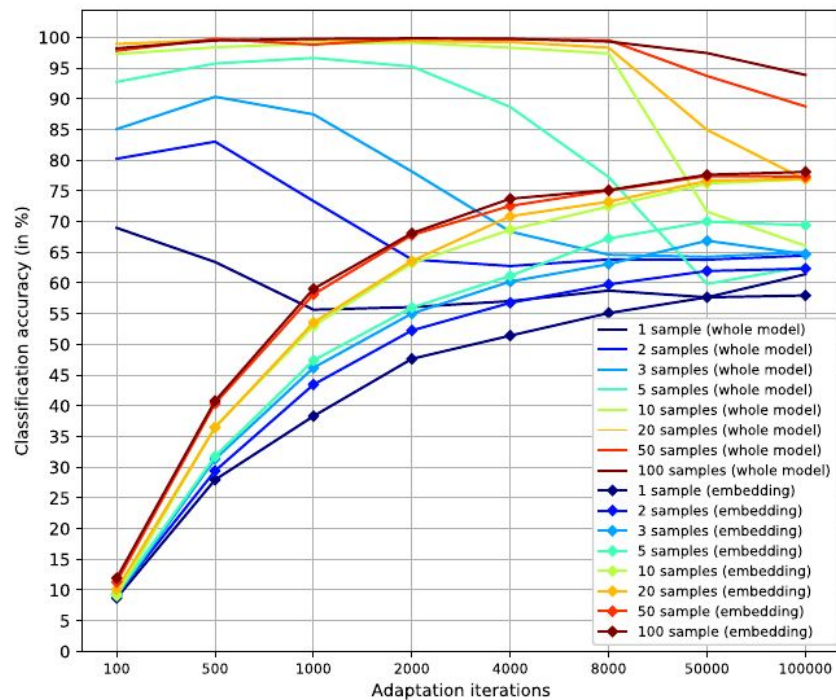
(a)



(b)

Figure 4: (a) Speaker classification accuracy with different numbers of cloning samples. (b) EER (using 5 enrollment audios) for different numbers of cloning samples. LibriSpeech (unseen speakers) and VCTK represent EERs estimated from random pairing of utterances from ground-truth datasets.

Whole model adaptation vs Speaker embedding adaptation



Naturalness

Approach	Sample count				
	1	2	3	5	10
Ground-truth (16 KHz sampling rate)	4.66±0.06				
Multi-speaker generative model	2.61±0.10				
Speaker adaptation (embedding-only)	2.27±0.10	2.38±0.10	2.43±0.10	2.46±0.09	2.67±0.10
Speaker adaptation (whole-model)	2.32±0.10	2.87±0.09	2.98±0.11	2.67±0.11	3.16±0.09
Speaker encoding (without fine-tuning)	2.76±0.10	2.76±0.09	2.78±0.10	2.75±0.10	2.79±0.10
Speaker encoding (with fine-tuning)	2.93±0.10	3.02±0.11	2.97±0.1	2.93±0.10	2.99±0.12

Table 2: Mean Opinion Score (MOS) evaluations for naturalness with 95% confidence intervals (training with LibriSpeech speakers and cloning with 108 VCTK speakers).

Similarity with ground-truth

Approach	Sample count				
	1	2	3	5	10
Ground-truth (same speaker)	3.91 ± 0.03				
Ground-truth (different speakers)	1.52 ± 0.09				
Speaker adaptation (embedding-only)	2.66 ± 0.09	2.64 ± 0.09	2.71 ± 0.09	2.78 ± 0.10	2.95 ± 0.09
Speaker adaptation (whole-model)	2.59 ± 0.09	2.95 ± 0.09	3.01 ± 0.10	3.07 ± 0.08	3.16 ± 0.08
Speaker encoding (without fine-tuning)	2.48 ± 0.10	2.73 ± 0.10	2.70 ± 0.11	2.81 ± 0.10	2.85 ± 0.10
Speaker encoding (with fine-tuning)	2.59 ± 0.12	2.67 ± 0.12	2.73 ± 0.13	2.77 ± 0.12	2.77 ± 0.11

Table 3: Similarity score evaluations with 95% confidence intervals (training with LibriSpeech speakers and cloning with 108 VCTK speakers).

Embedding manipulation

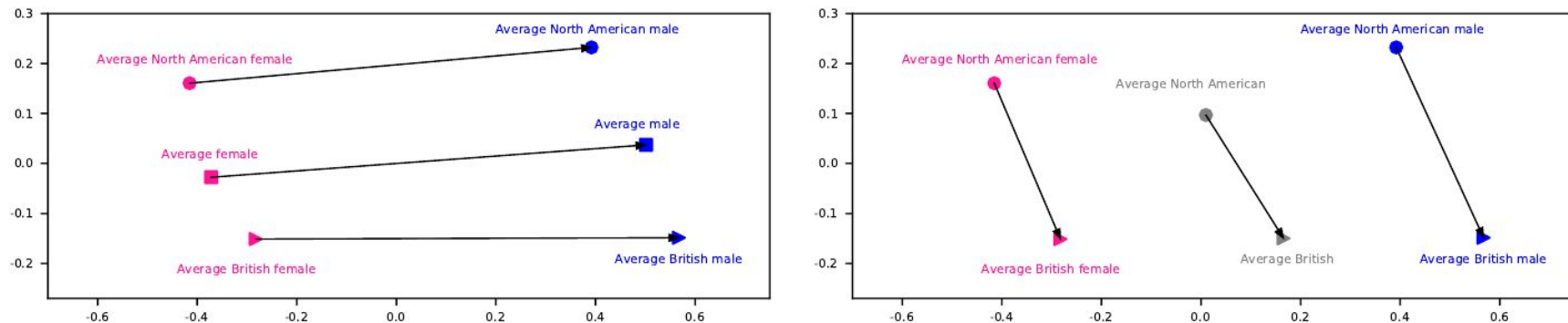


Figure 5: Visualization of estimated speaker embeddings by speaker encoder. The first two principal components of speaker embeddings (averaged across 5 samples for each speaker). Only British and North American regional accents are shown as they constitute the majority of the labeled speakers in the VCTK dataset. Please see Appendix E for more detailed analysis.