Neural Voice Cloning with a Few Samples

Sercan Ö. Arık, Jitong Chen*, Kainan Peng*, Wei Ping*, Yanqi Zhou*

** Equal Contribution*

Baidu Research

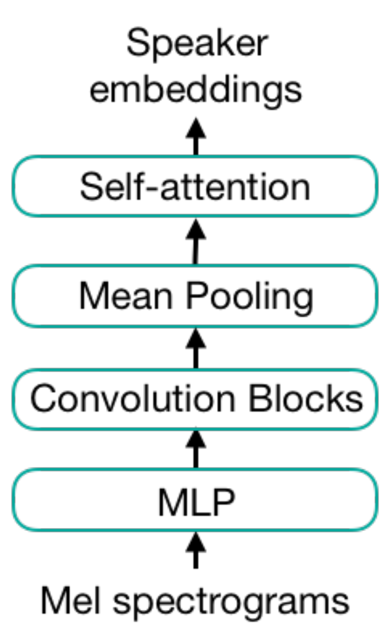
Summary

Introduction

- **Voice Cloning:** synthesize a person's voice from only a few audio samples, two approaches: speaker adaptation and speaker encoding.
- **Speaker Adaptation:** fine-tune a pre-trained multi-speaker speech synthesis model for an unseen speaker, two options: adaptation of speaker embedding only and adaptation of the whole model.
- **Speaker Encoding:** train a separate speaker encoder model to directly infer the speaker embedding for an unseen speaker, then feed the embedding into a multi-speaker speech synthesis model.

Architecture

- **Multi-speaker speech synthesis model:**
  - DeepVoice3, a fully convolutional sequence-to-sequence text-to-speech model training on more than 2000 speakers.
- **Speaker encoder model:**
  - *MLP:* spectral processing.
  - *Convolutional blocks and pooling*: temporal processing.
  - *Self-attention*: adaptively assign weights to different cloning audio samples while combining them.



Dataset

- For training multi-speaker speech synthesis model and speaker encoder: *LibriSpeech*: around 2500 speakers and 800 hours.
- For audio generation of speaker adaptation and speaker encoding: *VCTK*: 108 speakers around 40 hours.

Comparison

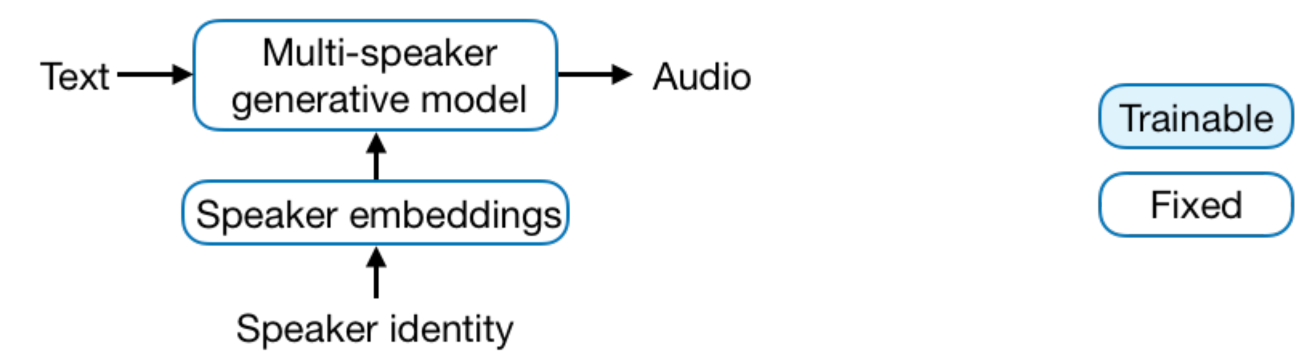
	Speaker Adaptation		Speaker Encoding	
Approaches	Embedding-only	Whole-model	Without fine-tuning	With fine-tuning
Training data	Text and audio		Audio and embedding	
Adaptation/Encoding time	8 h	0.5~5 min	1.5~3.5 s	1.5~3.5 s
Parameters per speaker	128	25 million	512	512

Conclusions

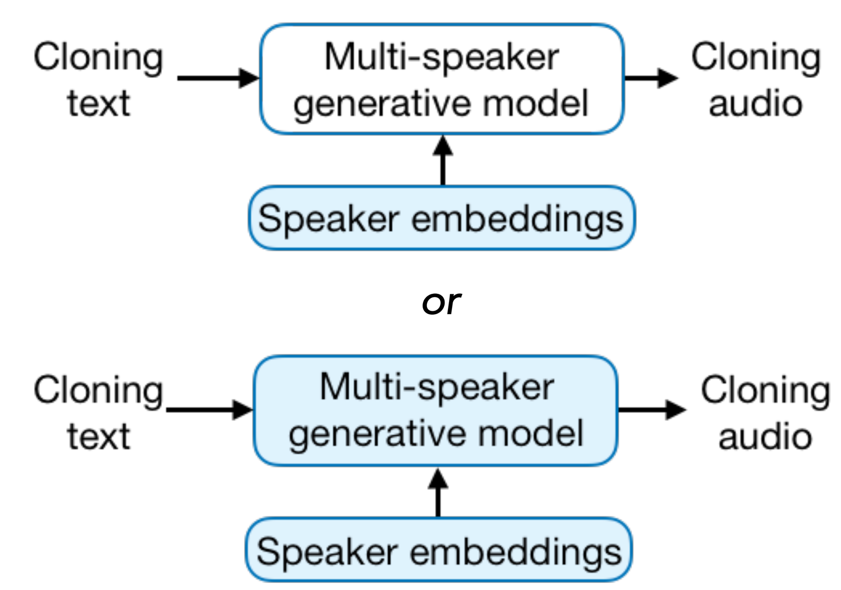
The adaptation/encoding time and required parameters per speaker are significantly less for speaker encoding approach, which makes it more favorable for low-resource deployment.

Neural Voice Cloning

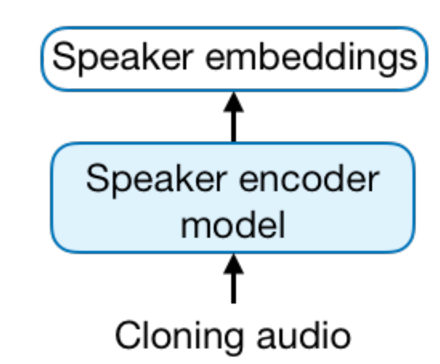
Pre-trained Multi-speaker Speech Synthesis Model



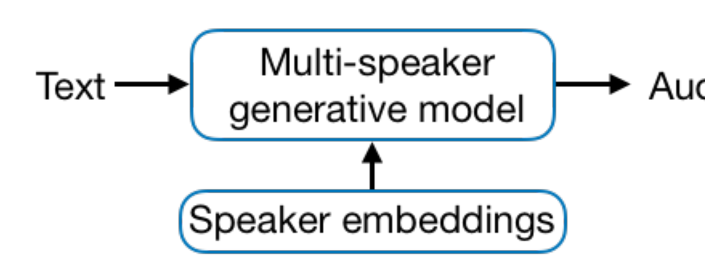
Speaker Adaptation Training

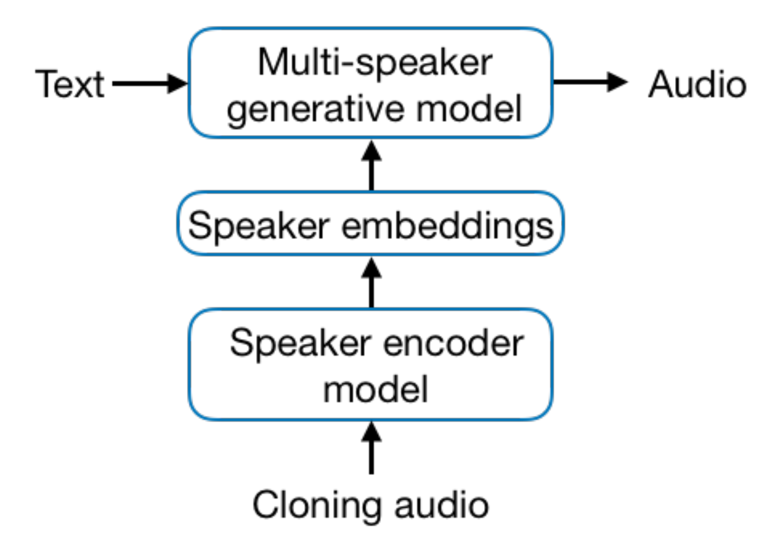


Speaker Encoding Training



Audio Synthesis

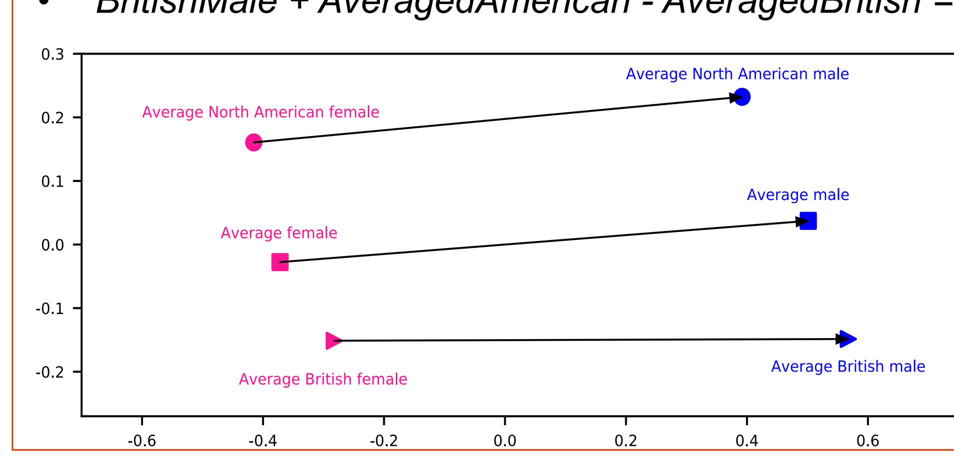


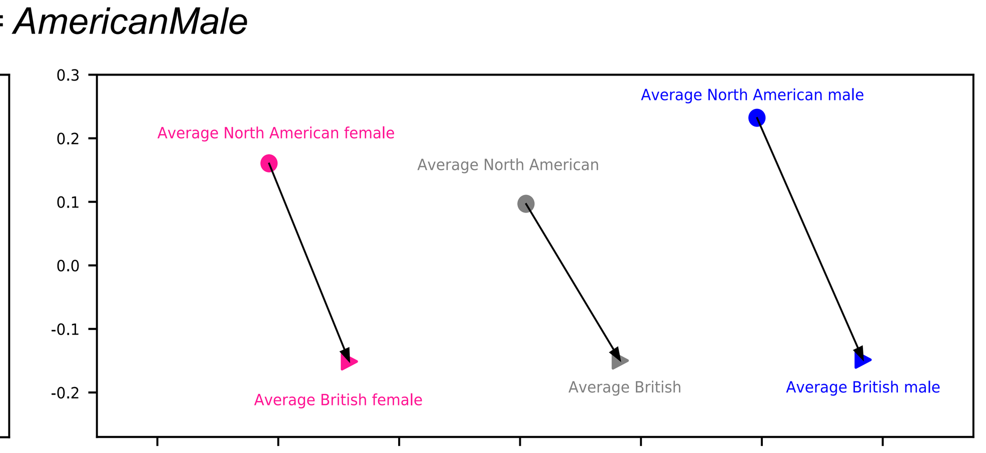


Voice Morphing via Embedding Manipulation

Simple algebraic operations to inferred embeddings are highly effective in transforming speaker characteristics.

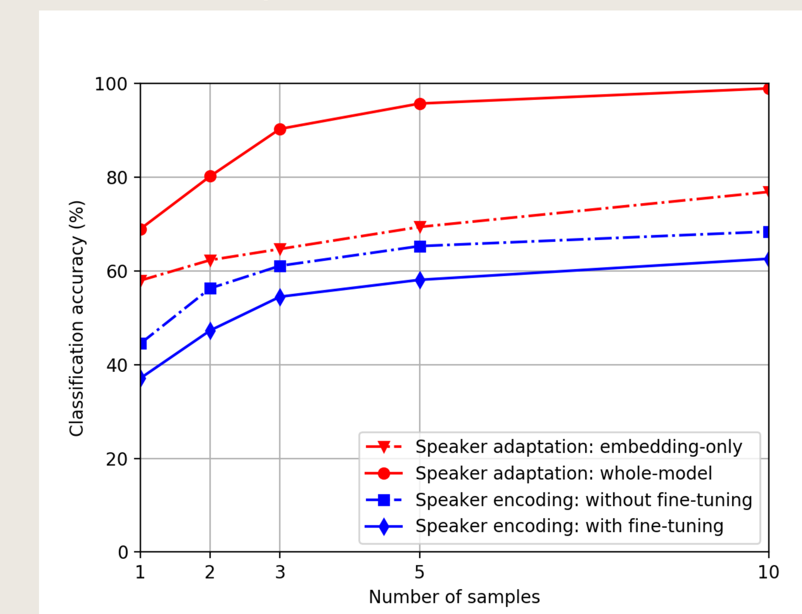
- *BritishMale + AveragedFemale - AveragedMale = BritishFemale*
- *BritishMale + AveragedAmerican - AveragedBritish = AmericanMale*



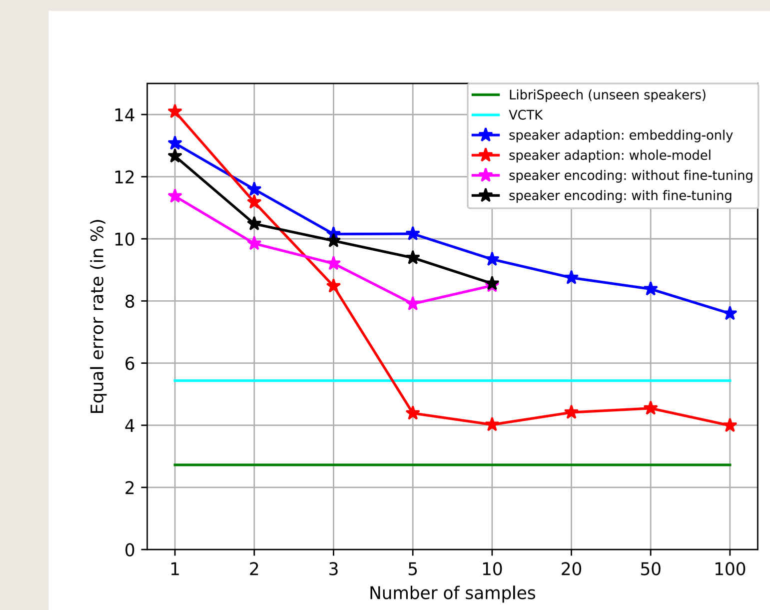


Automated Evaluation

Speaker Classification



Speaker Verification



Conclusions

- Both speaker adaptation and speaker encoding benefit from more cloning audios.
- When the number of cloning audio samples exceed five, whole-model adaptation outperforms others.
- Speaker encoding yields a lower classification accuracy compared to embedding adaptation, but it achieves a similar speaker verification performance.

Subjective Evaluation

Naturalness: 5-scale mean opinion score

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

Similarity: 4-scale similarity score

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

Conclusions

- Higher number of cloning audios improve both metrics. The improvement is more significant for whole model adaptation, due to the more degrees of freedom provided for an unseen speaker.
- Speaker encoding achieves naturalness similar or better than the baseline model. Similarity scores slightly improve with higher sample counts for speaker encoding, and match the scores for speaker embedding adaptation.