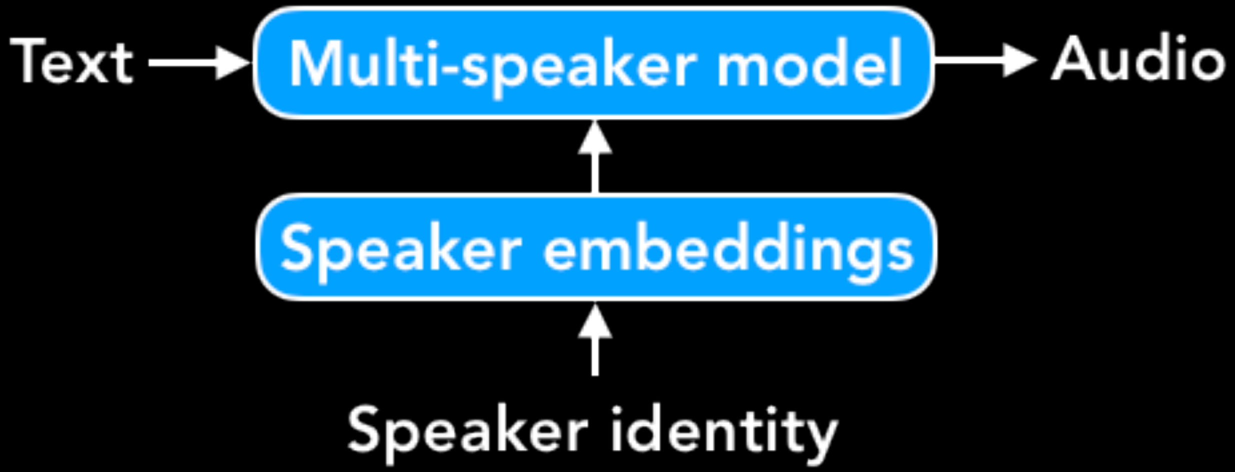
Neural Voice Cloning with a Few Samples

Sercan O. Arik, Jitong Chen, Kainan Peng*, Wei Ping, Yanqi Zhou

Baidu Research

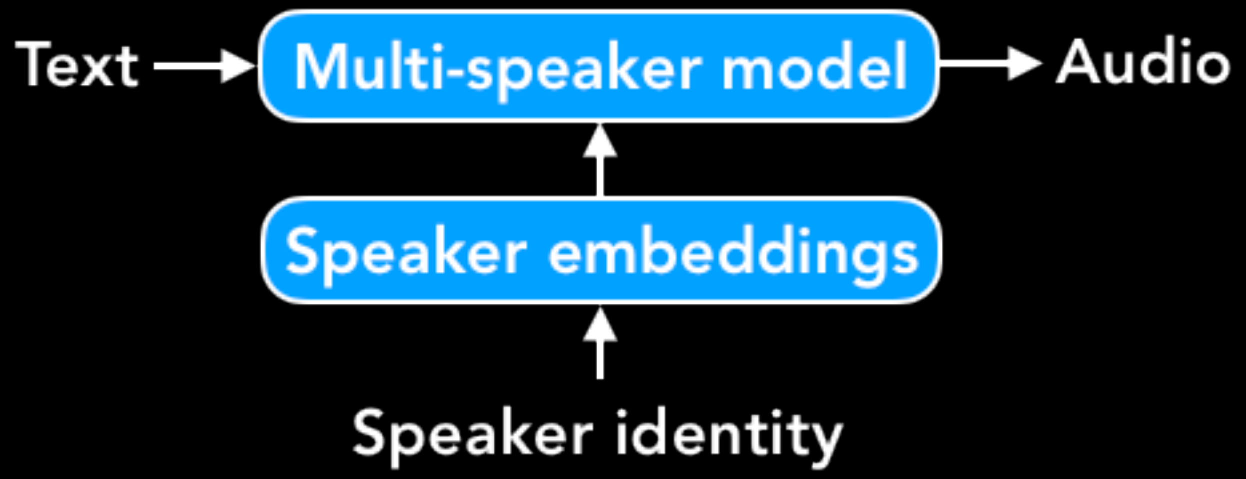
Motivations

- Text-to-speech (TTS) models can be conditioned on text and speaker identity.
  - Text: linguistic information, content of the generated speech.
  - Speaker identity: speaker information (accent, pitch, speech rate…).



Motivations

- Text-to-speech (TTS) models can be conditioned on text and speaker identity.
  - Text: linguistic information, content of the generated speech.
  - Speaker identity: speaker information (accent, pitch, speech rate…).



- Limitations:
  - Can only generate speech for observed speakers during training.
  - Require lots of speech samples per speaker (e.g., Deep Voice 2).

Voice Cloning

- Voice cloning: synthesize the voices of new speakers from a few speech samples (few-shot generative model).

- Applications: personalized speech interfaces, content creation, assistive technology…

Voice Cloning

- Voice cloning: synthesize the voices of new speakers from a few speech samples (few-shot generative model).

- Applications: personalized speech interfaces, content creation, assistive technology…

- Challenges:
  - Generalization: learn the voice of a new speaker.
  - Efficiency: extract the speaker characteristics from a few speech samples.
  - Computational cost: cloning with low latency and small footprint.

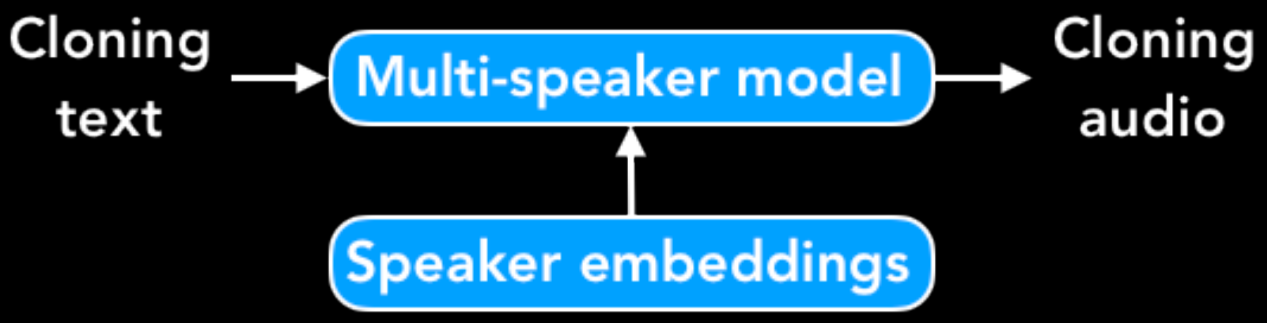
- Two approaches:
  - Speaker adaptation.
  - Speaker encoding.

Speaker Adaptation

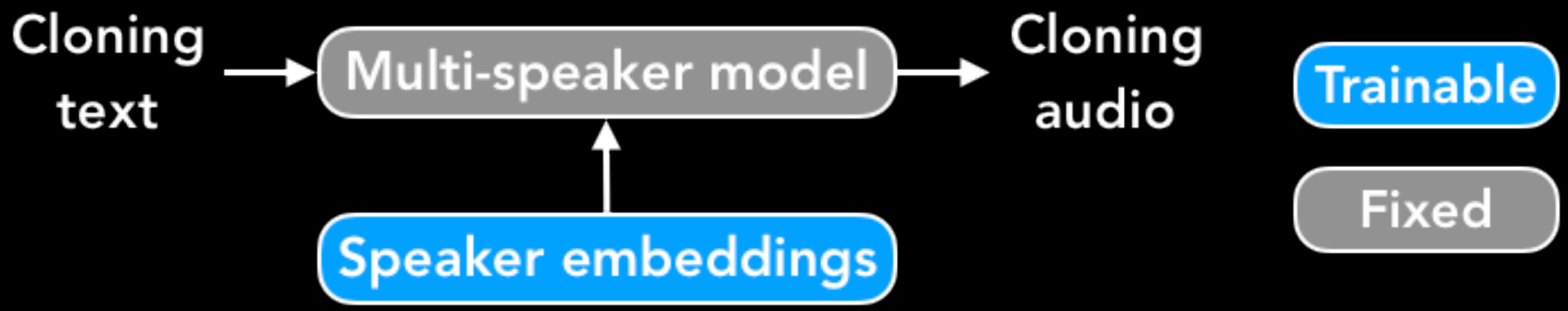
- Fine-tune a pre-trained multi-speaker model for a new speaker.
- Training data: a few text and audio pairs.

Speaker Adaptation

- Fine-tune a pre-trained multi-speaker model for a new speaker.
- Training data: a few text and audio pairs.
- Two options for speaker adaptation:



Fine-tune the whole model



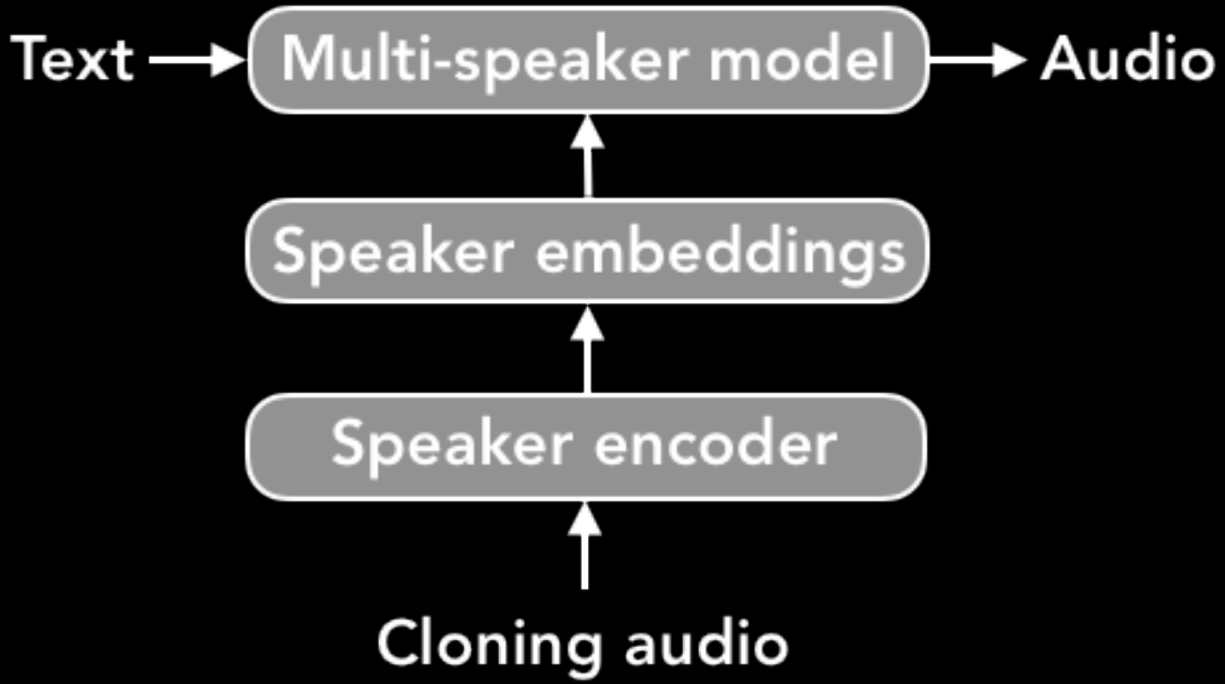
Fine-tune the speaker embedding only

Speaker Adaptation Analysis

Approaches	Speaker Adaptation	
	Embedding-only	Whole-model
Cloning time	8 h	5 min
# of parameters per speaker	128	25 million

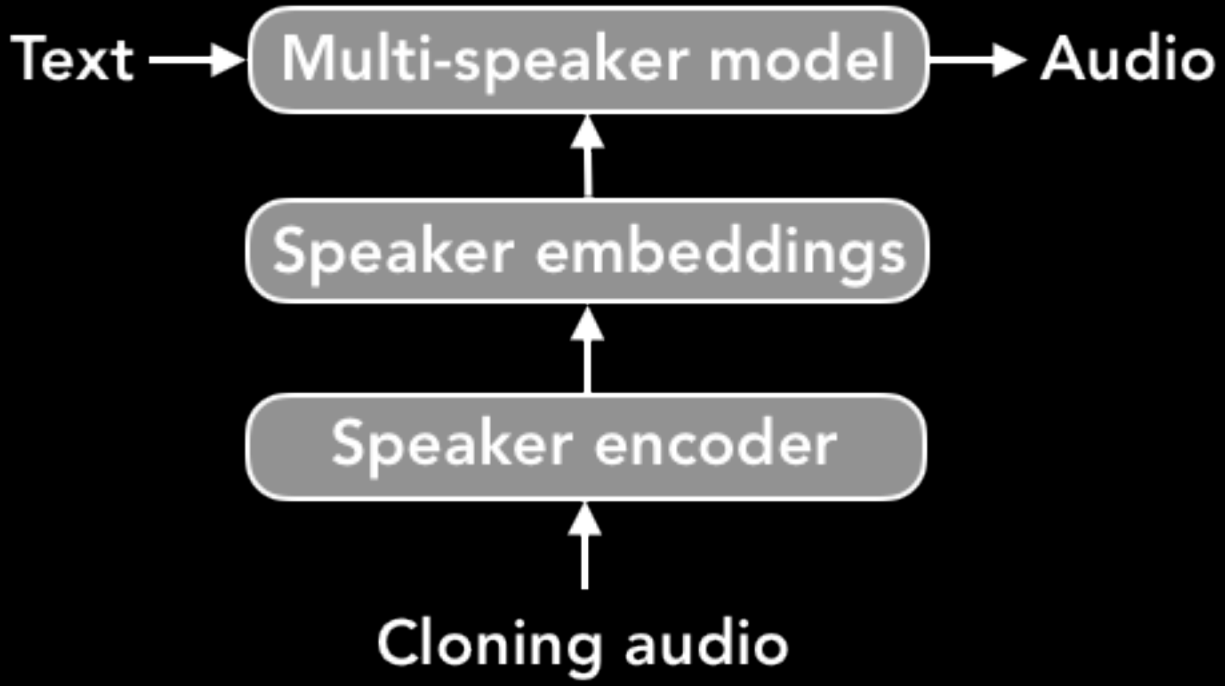
Speaker Encoding

- Directly predict a new speaker embedding for a multi-speaker model.
- Train a speaker encoder with audio and speaker embedding pairs.



Speaker Encoding

- Directly predict a new speaker embedding for a multi-speaker model.
- Train a speaker encoder with audio and speaker embedding pairs.
- Cloning time: a few seconds, more favorable for low-resource deployment.



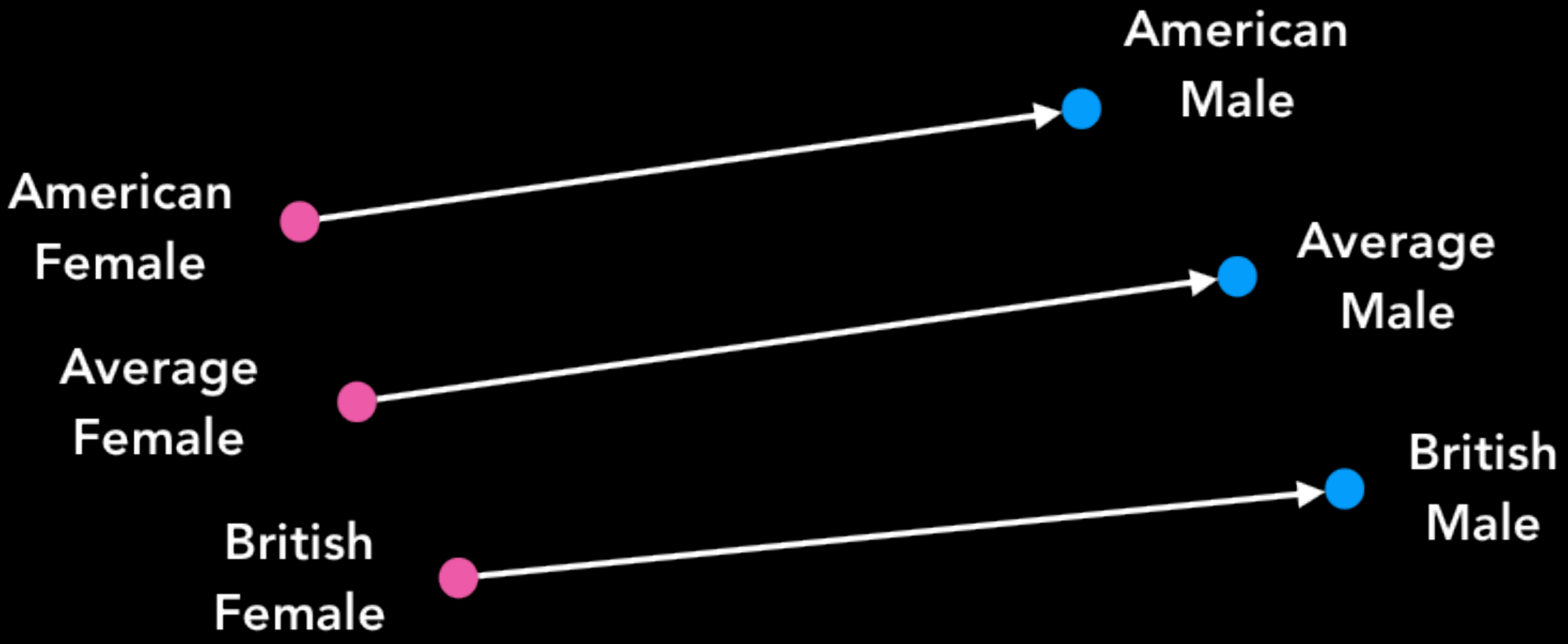
Results

- Vocoder: classical Griffin-Lim algorithm.
- Demo website: **http://audiodemos.github.io**

Approaches		Speaker Adaptation		Speaker Encoding
		Embedding-only	Whole-model	
Mean Opinion Score (MOS)	Naturalness (5-scale)	2.67	3.16	2.99
	Similarity (4-scale)	2.95	3.16	2.85

Voice Morphing via Embedding Manipulation

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



Thank you!

Welcome to our poster,
and listen to samples!

Today, Session B, #91

Baidu Research