Voice Clones with 5s duration of listening

transfer learning from speaker verification to multispeaker text-to-speech synthesis

Presenter: Yu Cheng-Hung
Thesis Advisor : Prof. Jian-Jiun Ding
Group meeting: 2022/05/24

AGENDA

01 Motivation

02 Introduction

03 Background

04 Model

05 Experiments

06 Conclusion

01

Motivation

Model tasks

1. More natural speech-to-speech translation
2. voice navigation
3. Address some safety concerns
4. Traditional TTS is data driven training
5. Transfer knowledge of speaker variability - speaker encoder

Reference speaker

Synthesized

Take a look at these pages for crooked creek drive

02

Introduction

Pitfall from previous research

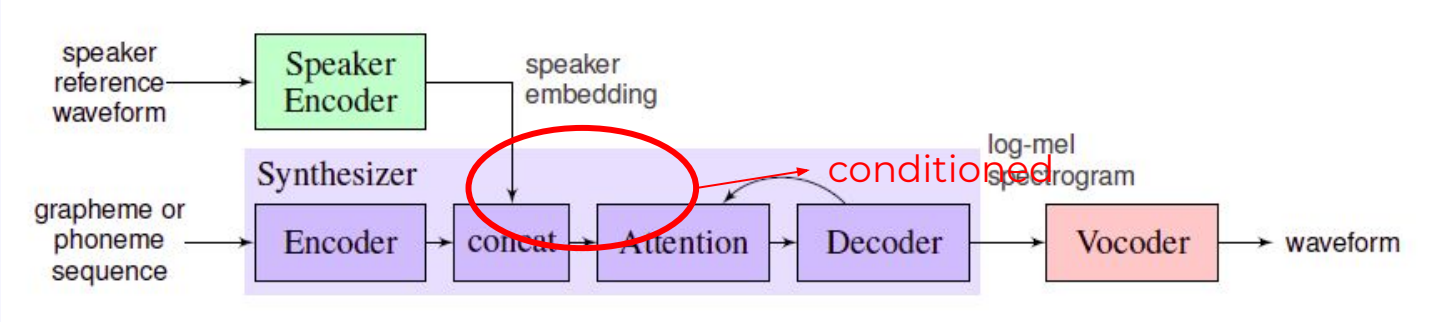
- Based on a large number of high quality speech-transcript pairs.
- Per speaker with 10 minutes of training data
- Speaker-dependent parameters are stored in a very low-dimensional vector
- Need several model to represent speaker

Found that simply speaker embedding is not enough for synthesis task and represent the mel spectrogram for wavenet

Ref: Deep Voice 2: Multi-speaker neural text-to-speech

Proposed approach

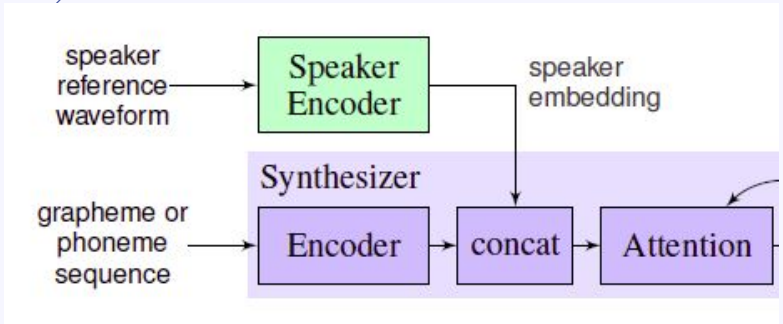
- decouple speaker modeling from speech synthesis - captures the space of speaker characteristics
- Training TTS model conditioned on speaker encoder



REF: Transfer learning from speaker verification to multispeaker text-to-speech synthesis.

Speaker encoder

Deep Voice 2: Multi-speaker neural text-to-speech



Previous research :

High quality multi-speaker training data: Usually, each speaker has its own embedding table

Improvement

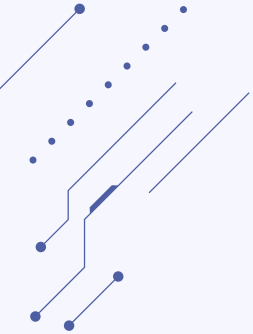
1. **Reducing the need of HQ data**

Trained on thousands of speakers and squeezed all the trained data into embedding space.

2. **Denoise and dereverberation**

The encoder learns the essence of human speech from many speakers and noise is not essence of human speech.

Unseen speaker and seen speaker



Previous research :

1. **Only learn a fixed set of speaker embeddings table**
   Each speaker has its own training speaker.
2. **Support only synthesis of voices seen during training**
   But some apply predicting mechanism on speaker embedding for unseen voice.

Improvement

1. **Train the encoder on task of discriminate between speaker**
   This helps encoder to learn the transfer of speaker characteristics.
2. **Use combination of embedding space to generate unseen speaker**

03

Background

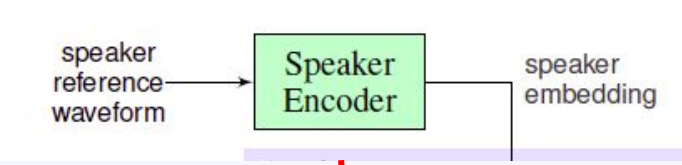
System with three components

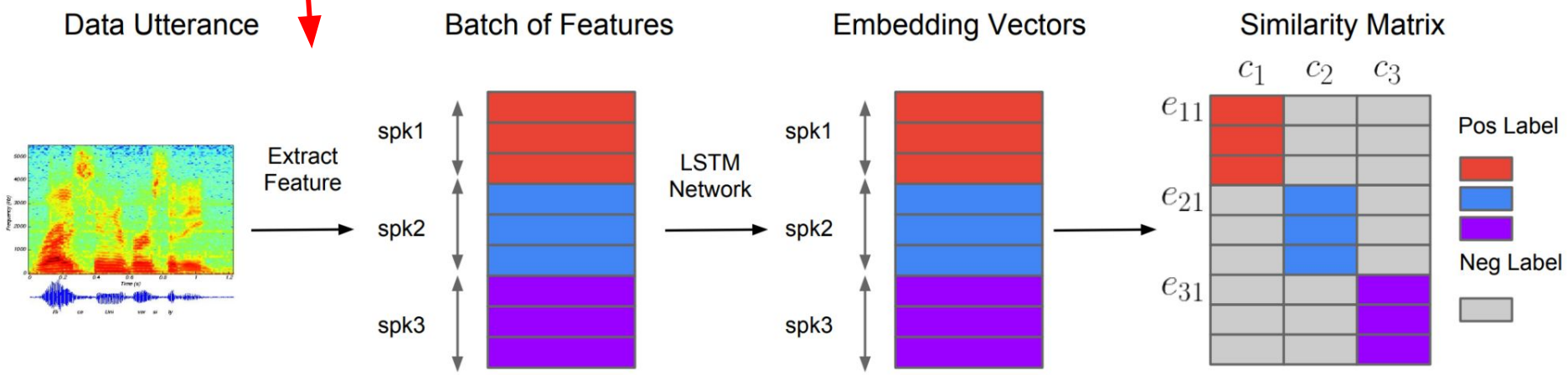
Speaker encoder network

Sequence to sequence synthesis network

Auto-regressive Wavenet-based vocoder

Recurrent speaker encoder

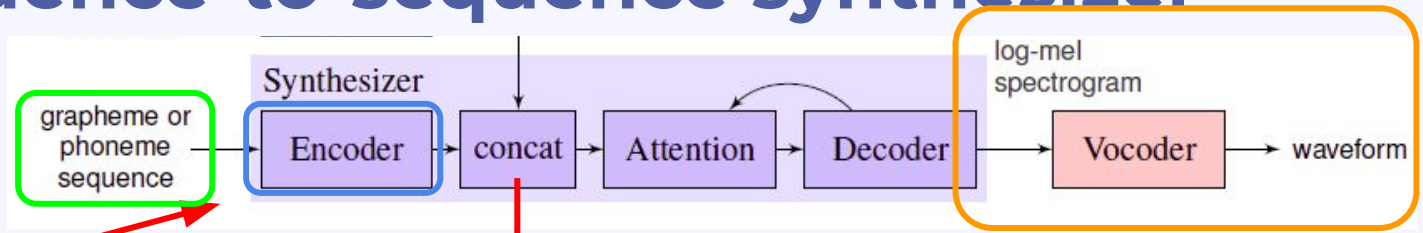


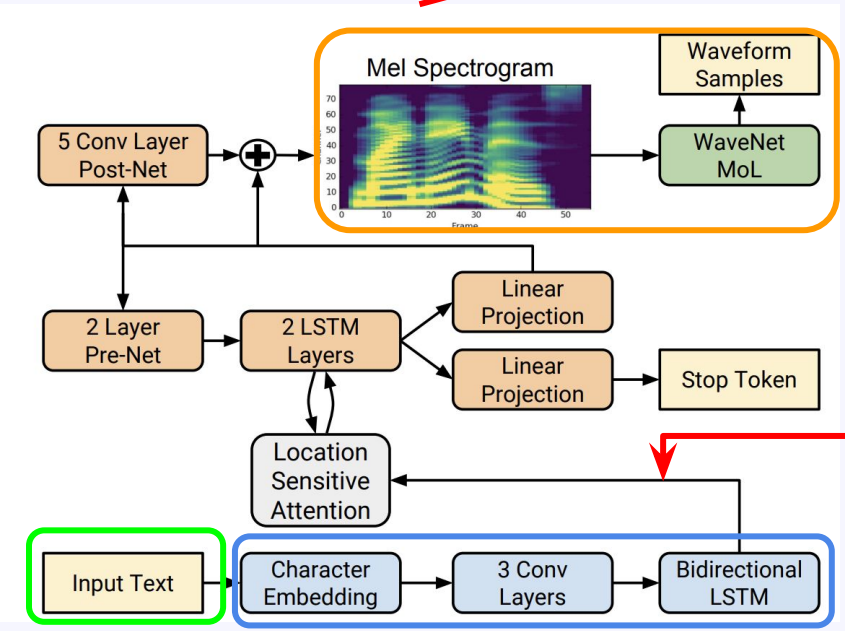


Ref: GENERALIZED END-TO-END LOSS FOR SPEAKER VERIFICATION

Fig. 1. System overview. Different colors indicate utterances/embeddings from different speakers.

Sequence-to-sequence synthesizer



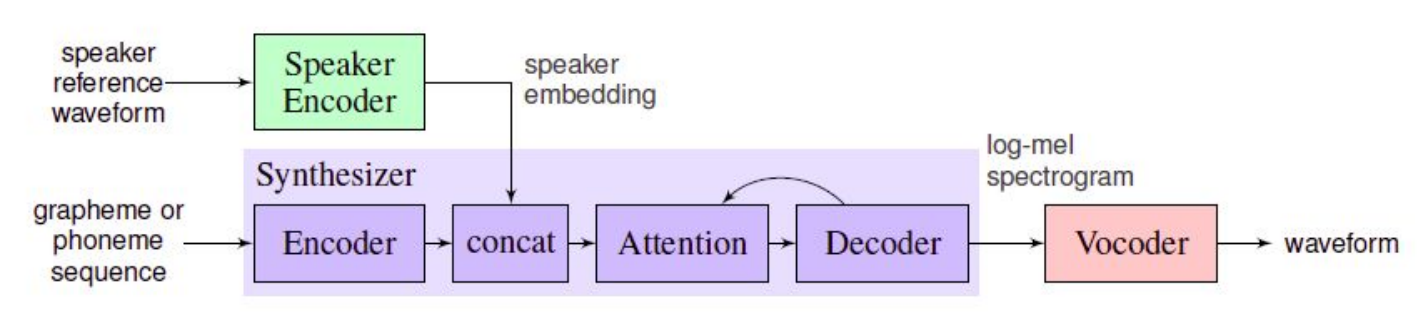


Ref: Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions.

04

Model

Model structure



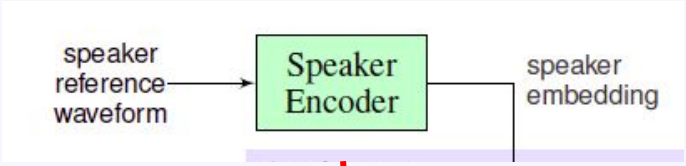
Speaker Encoder
- Tries to learn the essence of human speech from thousands of speakers

Synthesizer
- Clone the test speaker with input text

Vocoder
- Transfer the mel spectrogram with embedding text to audio waveform

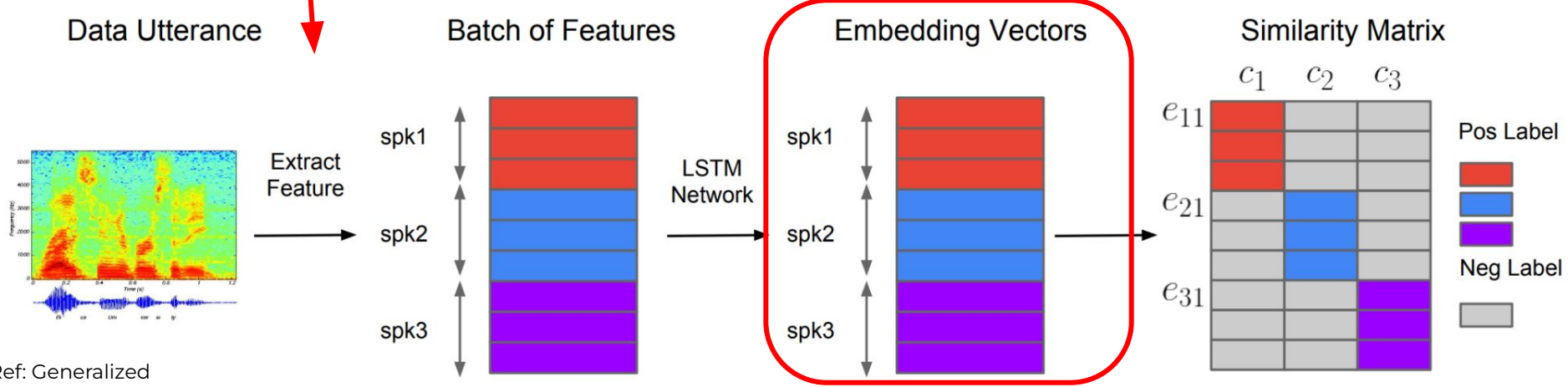
REF: Transfer learning from speaker verification to multispeaker text-to-speech synthesis.

Recurrent speaker encoder



$$\mathbf{S}_{ji,k} = w \cdot \cos(\mathbf{e}_{ji}, \mathbf{c}_k) + b,$$

D-vector
- fixed dimensional vector from a speech signal



Ref: Generalized end-to-end loss for speaker verification.

Fig. 1. System overview. Different colors indicate utterances/embeddings from different speakers.

speaker encoder verification task

Ref: Generalized end-to-end loss for speaker verification.

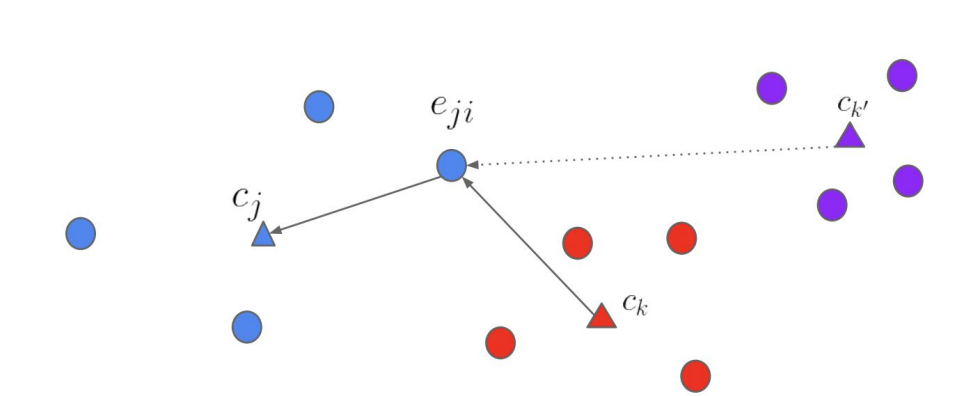


Fig. 2. GE2E loss pushes the embedding towards the centroid of the true speaker, and away from the centroid of the most similar different speaker.

softmax loss

$$L(\mathbf{e}_{ji}) = -\mathbf{S}_{ji,j} + \log \sum_{k=1}^{N} \exp(\mathbf{S}_{ji,k}).$$

Positive pair distance

Negative pair distance

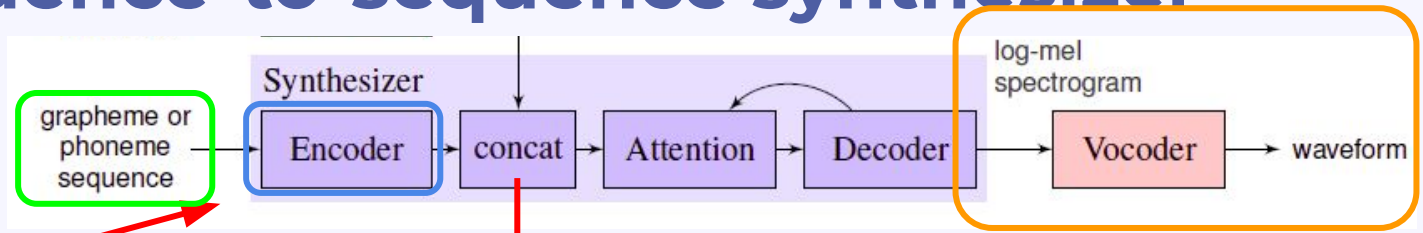
contrast loss

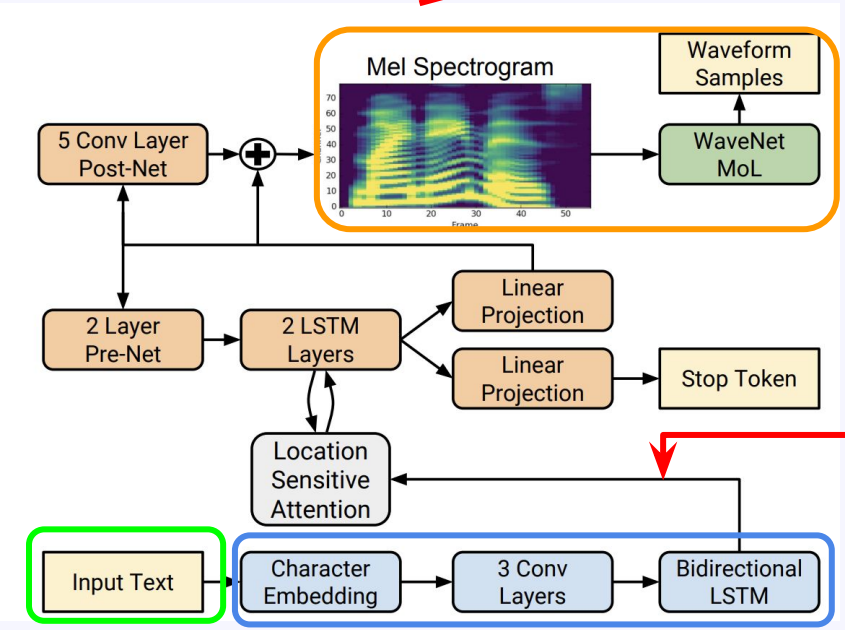
$$L(\mathbf{e}_{ji}) = 1 - \sigma(\mathbf{S}_{ji,j}) + \max_{\substack{1 \le k \le N \\ k \ne j}} \sigma(\mathbf{S}_{ji,k}),$$

Minimize final loss

$$L_G(\mathbf{x}; \mathbf{w}) = L_G(\mathbf{S}) = \sum_{j,i} L(\mathbf{e}_{ji}).$$

Sequence-to-sequence synthesizer



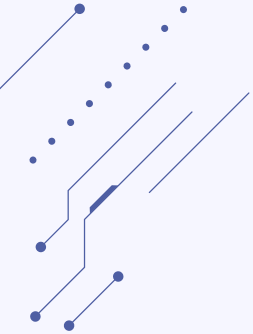


Input: LEX -> phonemes: L EH1 K S

Ref: Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions.

Inference and zero-shot speaker adaptation

- **Encoder captures the essence of speaker**
  Can still synthesis and conditioned on unseen audio
- **Zero-shot adaptation to novel speaker**
  After one training, single audio clip of 5s is enough for synthesis



Synthesis in spectrogram

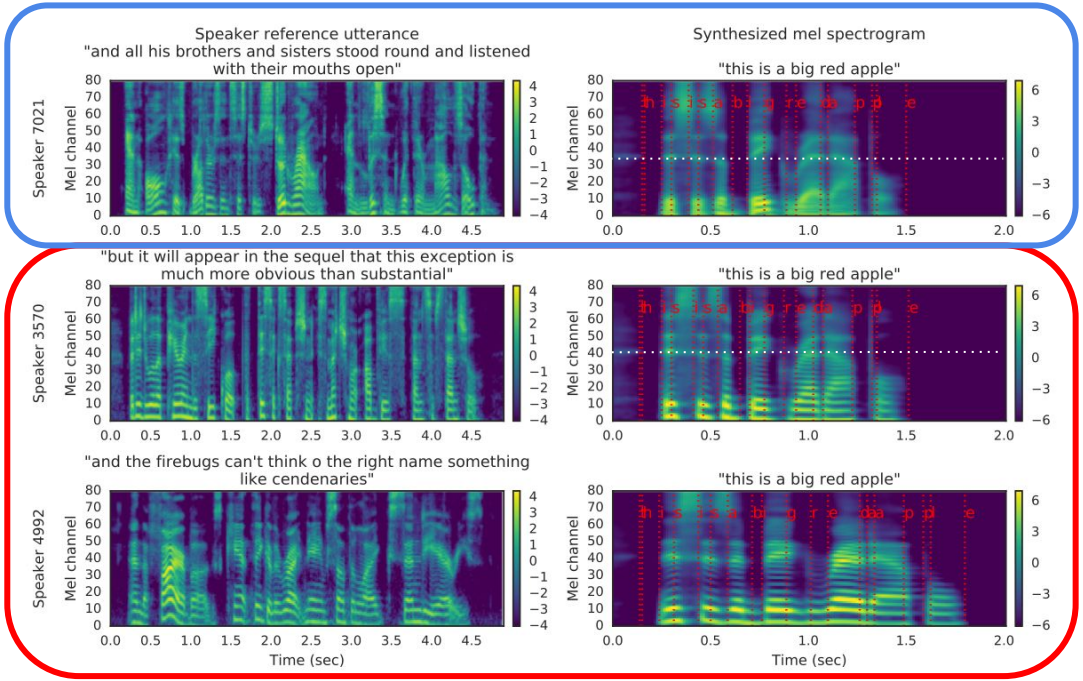


Figure 2: Example synthesis of a sentence in different voices using the proposed system. Mel spectrograms are visualized for reference utterances used to generate speaker embeddings (left), and the corresponding synthesizer outputs (right). The text-to-spectrogram alignment is shown in red. Three speakers held out of the train sets are used: one male (top) and two female (center and bottom).

Top has noticeably lower F0

0.3s
Same “i” sound, male has Mel 35
But female has Mel 40

0.4s
Same phenomenon with “s” sound

Encoder also learn the speaker rate

REF: Transfer learning from speaker verification to multispeaker text-to-speech synthesis.

05

Experiments

Speech naturalness

Table 1: Speech naturalness Mean Opinion Score (MOS) with 95% confidence intervals.

System	VCTK Seen	VCTK Unseen	LibriSpeech Seen	LibriSpeech Unseen
Ground truth	4.43 ± 0.05	4.49 ± 0.05	4.49 ± 0.05	4.42 ± 0.07
Embedding table	4.12 ± 0.06	N/A	3.90 ± 0.06	N/A
Proposed model	4.07 ± 0.06	4.20 ± 0.06	3.89 ± 0.06	4.12 ± 0.05

LibriSpeech: noisy data with US accent

- Encoder train on noisy data
- Synthesis train and use on preprocessed data

VCTK: cleaner data with British accent

Bad	Description
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

REF: Transfer learning from speaker verification to multispeaker text-to-speech synthesis.

Speaker similarity

Table 2: Speaker similarity Mean Opinion Score (MOS) with 95% confidence intervals.

System	Speaker Set	VCTK	LibriSpeech
Ground truth	Same speaker	4.67 ± 0.04	4.33 ± 0.08
Ground truth	Same gender	2.25 ± 0.07	1.83 ± 0.07
Ground truth	Different gender	1.15 ± 0.04	1.04 ± 0.03
Embedding table	Seen	4.17 ± 0.06	3.70 ± 0.08
Proposed model	Seen	4.22 ± 0.06	3.28 ± 0.08
Proposed model	Unseen	3.28 ± 0.07	3.03 ± 0.09

REF: Transfer learning from speaker verification to multispeaker text-to-speech synthesis.

Table 3: Cross-dataset evaluation on naturalness and speaker similarity for unseen speakers.

Synthesizer Training Set	Testing Set	Naturalness	Similarity
VCTK	LibriSpeech	4.28 ± 0.05	1.82 ± 0.08
LibriSpeech	VCTK	4.01 ± 0.06	2.77 ± 0.08

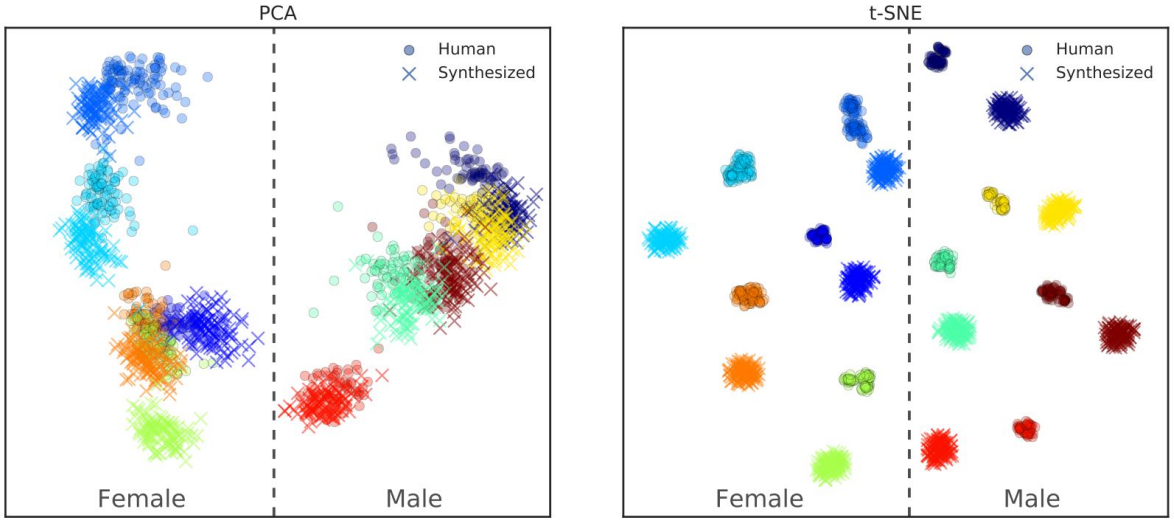


Figure 3: Visualization of speaker embeddings extracted from LibriSpeech utterances. Each color corresponds to a different speaker. Real and synthetic utterances appear nearby when they are from the same speaker, however real and synthetic utterances consistently form distinct clusters.

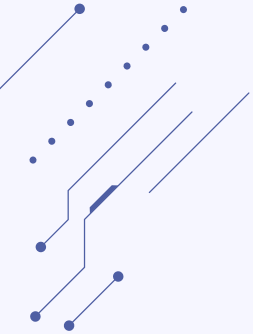
REF: Transfer learning from speaker verification to multispeaker text-to-speech synthesis.

06

Conclusion

Conclusion

1. Reducing the speaker data need
2. Able to synthesis unseen speaker
3. The model occasionally learn the speaker rate
4. The encoder can learn and mimic the human voice



07

Reference

Jia, Ye, et al. "Transfer learning from speaker verification to multispeaker text-to-speech synthesis." Advances in neural information processing systems 31 (2018).

Wan, Li, et al. "Generalized end-to-end loss for speaker verification." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

Gibiansky, Andrew, et al. "Deep voice 2: Multi-speaker neural text-to-speech." Advances in neural information processing systems 30 (2017).

Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018.

Thank you !