Sample Efficient Adaptive Text-to-Speech

Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C. Cobo, Andrew Trask, Ben Laurie, Caglar Gulcehre, Aäron van den Oord, Oriol Vinyals, Nando de Freitas

DeepMind & Google

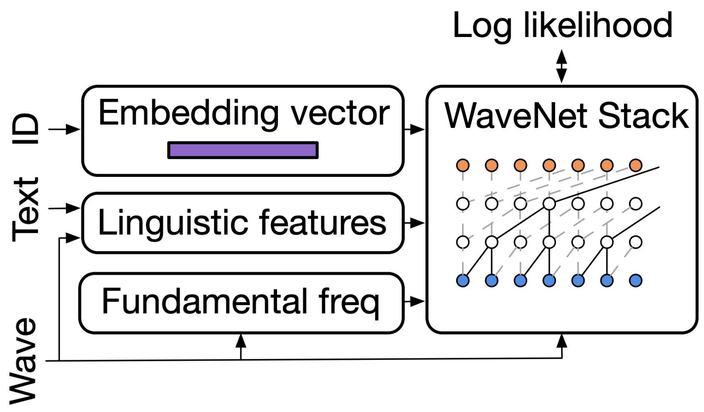
Demo

Highlights

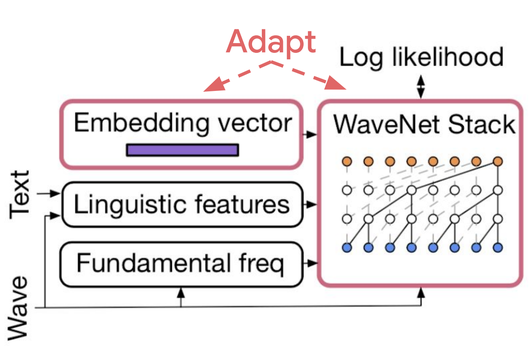
- Solving many tasks with few data, as opposed to solving few tasks with many data.
- Few-shot meta-learning enables us to **imitate a new voice-style with 5 minutes of data as opposed to the previous requirement of tens of hours**, and without compromising on quality.
- Our models contain both task-dependent and task-independent cores. This division facilitates training and enables fast adaptation to novel voices with low sample complexity.
- We achieve **state-of-the-art in both sample *naturalness* and *voice similarity*** with merely a few minutes of audio data.

Training and Adaptation

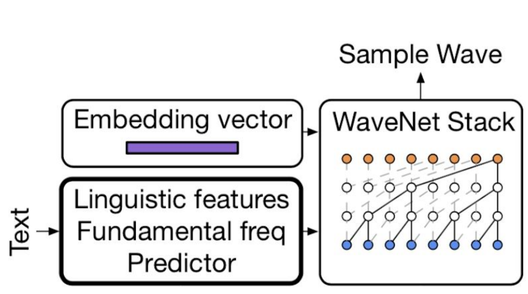
The adaptation consists of three stages:



a) Training: Train a multi-speaker model with a shared WaveNet core and independent learned embeddings for each speaker (task)



b) Adaptation: Adapt to a new speaker with few-shot data



c) Inference: Generate new speech with the adapted model

Experimental Procedure

Train:
- More than 2300+ speakers.
- 300-hour of Google American English TTS corpus.
- 500-hour public LibriSpeech corpus.

Adapt:
- Adapt on a hold-out speaker for:
- LibriSpeech: 10 sec - 5 min; VCTK: 10 sec - 10 min

Evaluate:
- (Subjective) Mean Opinion Score (MOS) human ratings:
  - for naturalness, and
  - for speaker similarity.
- (Objective) Speaker embedding (*d*-vectors) similarity from Google's Text-independent Speaker Verification model (Wan et al., 2018).

Naturalness (Mean Opinion Score)

Subjects were asked to rate the naturalness of generated utterances on a five-point Likert Scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent).

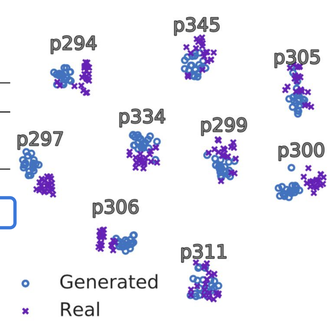
Dataset	LibriSpeech		VCTK	
Real utterance	4.38 ± 0.04		4.45 ± 0.04	
van den Oord et al. (2016)	4.21 ± 0.081			
Nachmani et al. (2018)	2.53 ± 1.11		3.66 ± 0.84	
Arik et al. (2018)				
adapt embedding	-		2.67 ± 0.10	
adapt whole-model	-		3.16 ± 0.09	
encoding + fine-tuning	-		2.99 ± 0.12	
Jia et al. (2018)				
trained on LibriSpeech	$\mathbf{4.12 \pm 0.05}$		$\mathbf{4.01 \pm 0.06}$	
Adaptation data size	*10s*	*<5m*	*10s*	*<10m*
SEA-ALL (ours)	3.94 ± 0.08	$\mathbf{4.13 \pm 0.06}$	$\mathbf{3.92 \pm 0.07}$	$\mathbf{3.92 \pm 0.07}$
SEA-EMB (ours)	3.86 ± 0.07	3.95 ± 0.07	3.81 ± 0.07	3.82 ± 0.07
SEA-ENC (ours)	3.61 ± 0.06	3.56 ± 0.06	3.65 ± 0.06	3.58 ± 0.06

Voice Similarity (Mean Opinion Score and d-vectors)

(MOS) Subjects were asked to rate the similarity between real and synthesized samples of the same speaker. (1: Not at all similar, 2: Slightly similar, 3: Moderately similar, 4: Very similar, 5: Extremely similar)

Dataset	LibriSpeech		VCTK	
Real utterance	4.30 ± 0.08		4.59 ± 0.06	
Jia et al. (2018)				
trained on LibriSpeech	3.03 ± 0.09		2.77 ± 0.08	
Adaptation data size	*10s*	*<5m*	*10s*	*<10m*
SEA-ALL (ours)	$\mathbf{3.41 \pm 0.10}$	$\mathbf{3.75 \pm 0.09}$	$\mathbf{3.51 \pm 0.10}$	$\mathbf{3.97 \pm 0.09}$
SEA-EMB (ours)	3.42 ± 0.10	3.56 ± 0.10	3.07 ± 0.10	3.18 ± 0.10
SEA-ENC (ours)	2.47 ± 0.09	2.59 ± 0.09	2.07 ± 0.08	2.19 ± 0.09

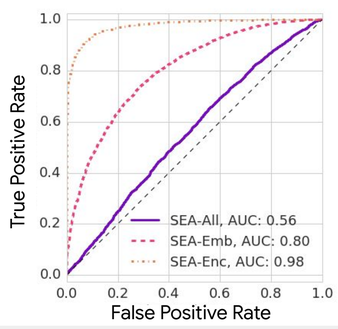
Real vs Synthesized d-vectors (t-SNE)



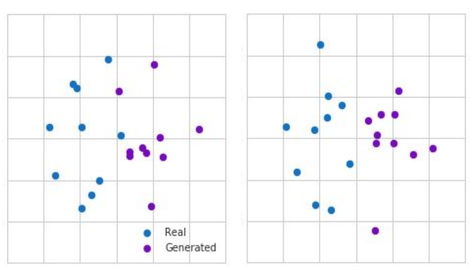
Detecting Synthesized Samples

- Preliminary results show that a linear SVM trained on the d-vectors of real and synthesized multi-speaker samples from our model achieves 85% accuracy.
- Anti-spoofing is an important area for continued research.

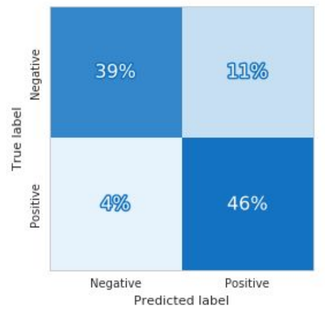
Standard speaker verification model failed detecting synthesized utterances



Observation: partially linearly separable on single speaker d-vectors

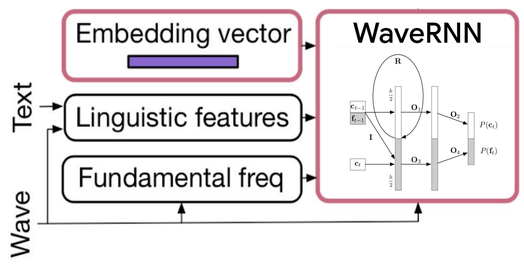


Linear SVM on d-vectors (85% accuracy)



Beyond the Paper: Faster Inference Models

- It is possible to use a WaveRNN model to achieve faster inference, without compromising on sample quality.
- Few-shot adaptation for WaveRNN works out-of-the-box.



Conclusions

- Impressive performance even with only 10 seconds of audio from new speakers.
- State-of-the-art performance in naturalness and voice similarity.
- Applicable to restoring the voices of speech-impaired patients.
- Promising results on detection of synthesized voices.