Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

Ye Jia Yu Zhang Ron J. Weiss Quan Wang Jonathan Shen Fei Ren Zhifeng Chen Patrick Nguyen Ruoming Pang Ignacio Lopez Moreno Yonghui Wu

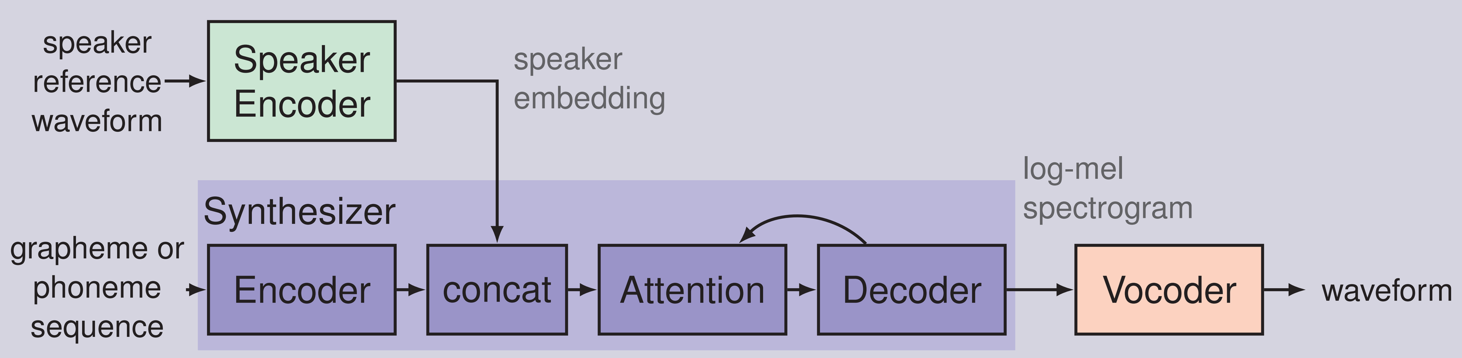
Google

{jiaye,ngyuzh,ronw}@google.com

1. Summary

- Multispeaker Tacotron 2 TTS network conditioned on *speaker embedding* computed from reference utterance
- Similar to [1, 2] except we focus on transfer from **pretrained** speaker encoder
- Make efficient use of available training data
  - untranscribed and noisy audio to train speaker encoder
  - smaller dataset of clean speech to train synthesizer
- Generalizes better than joint training only using clean speech dataset
- Allows zero-shot adaptation from ∼5 second reference utterance
  - although result is still distinguishable from real speech from that speaker
- Performance improves with number of speaker encoder training speakers

2. System architecture



1. Speaker encoder computes speaker embedding from spectrogram
   - stacked LSTM with 3 layers, embedding taken from output at final frame
   - discriminatively trained on speaker verification task [5]
2. Synthesizer generates mel spectrogram from input phoneme sequence
   - sequence-to-sequence model with attention, based on Tacotron 2 [3]
3. Vocoder inverts spectrogram to time-domain waveform
   - conditional WaveNet [4], 30 dilated convolution layers

3. Experiments

- Train speaker encoder on internal corpus of 39K hours from 18K speakers
  - noisy and reverberant speech without transcripts
- Train synthesizers and vocoders on clean, read speech from
  LibriSpeech (clean subset) 436 hours from ∼1.2K speakers
  VCTK 44 hours from 109 mostly British speakers
  - hold out 10 speakers from training to evaluate adaptation to unseen speakers
- Metrics
  - Subjective mean opinion score ratings of *speech naturalness* (MOS-nat) and *speaker similarity* (MOS-sim)
  - Speaker verification equal error rate (SV-EER), measured using eval-only speaker encoder trained on separate dataset

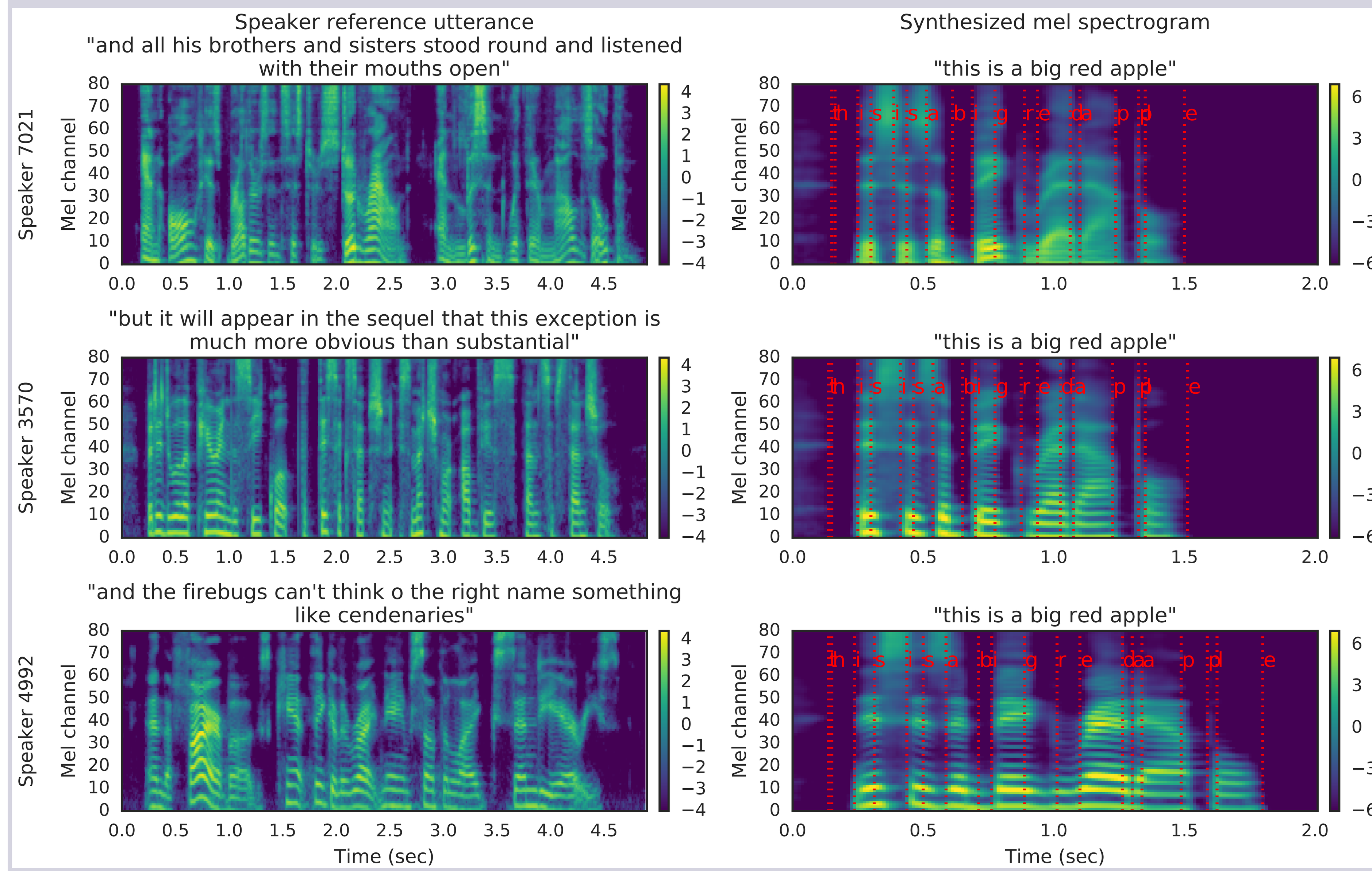
4. Results

		Train on VCTK			Train on LibriSpeech		
System	Speaker set	MOS-nat	MOS-sim	SV-EER	MOS-nat	MOS-sim	SV-EER
Ground truth	Seen	4.43	–	–	4.49	–	–
Ground truth	Unseen	4.49	4.67	1.5%	4.42	4.33	0.9%
Lookup table	Seen	4.12	4.17	1.2%	3.90	3.70	3.1%
Proposed	Seen	4.07	4.22	1.6%	3.89	3.28	4.3%
Proposed	Unseen	4.20	3.28	10.5%	4.12	3.03	5.1%
Proposed	Cross dataset	4.28	1.82	29.2%	4.01	2.77	6.3%

- Proposed has similar performance to lookup-table baseline on seen speakers
  - LibriSpeech generally has worse performance than VCTK
- Speaker similarity decreases (SV-EER increases) on unseen speakers
  - but much smaller change for LibriSpeech than VCTK
- Model trained on LibriSpeech can generalize to VCTK speakers
  - but does not transfer accents

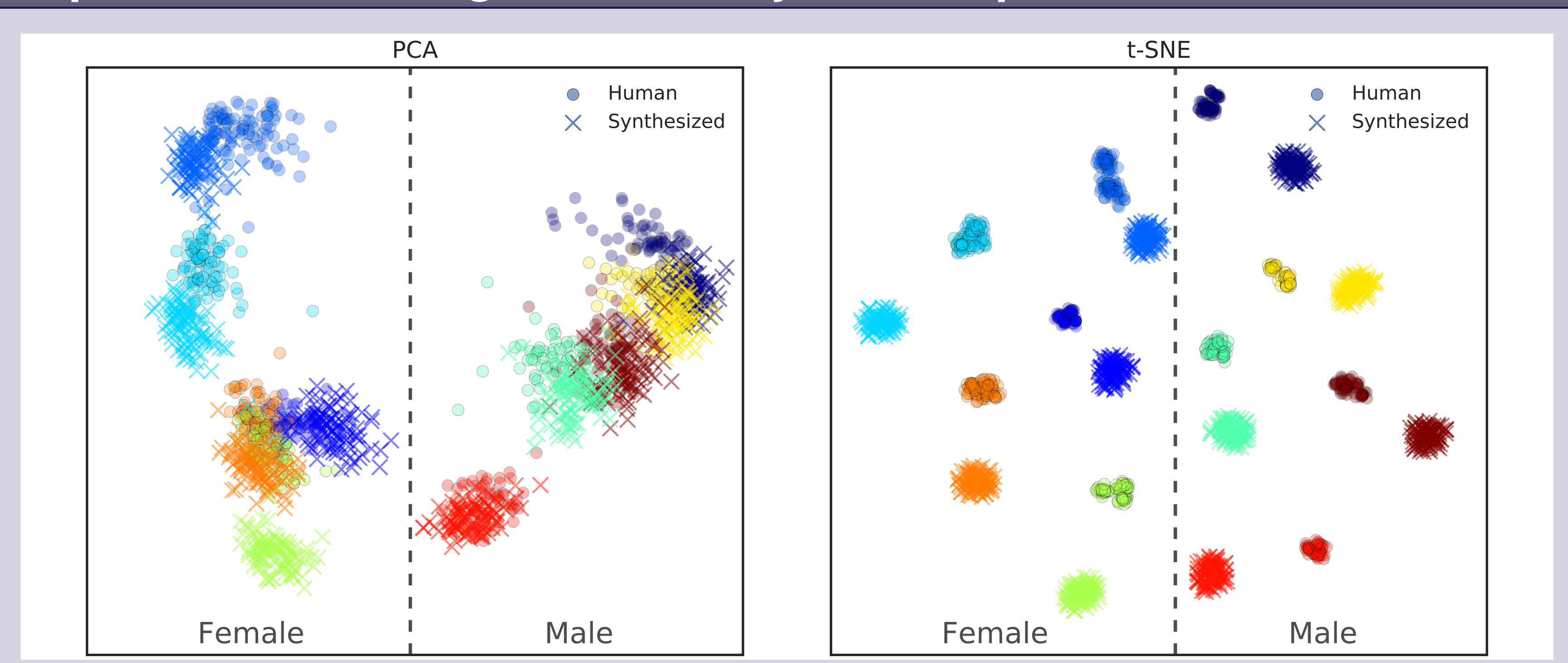
⟹ need to train synthesizer on many speakers

5. Synthesis examples



- Synthesize text using reference utterances from male (top), female (middle, bottom) speakers
- Different speaking rates and pitch/formant ranges, matching reference

6. Speaker embeddings: Real vs Synthetic speakers

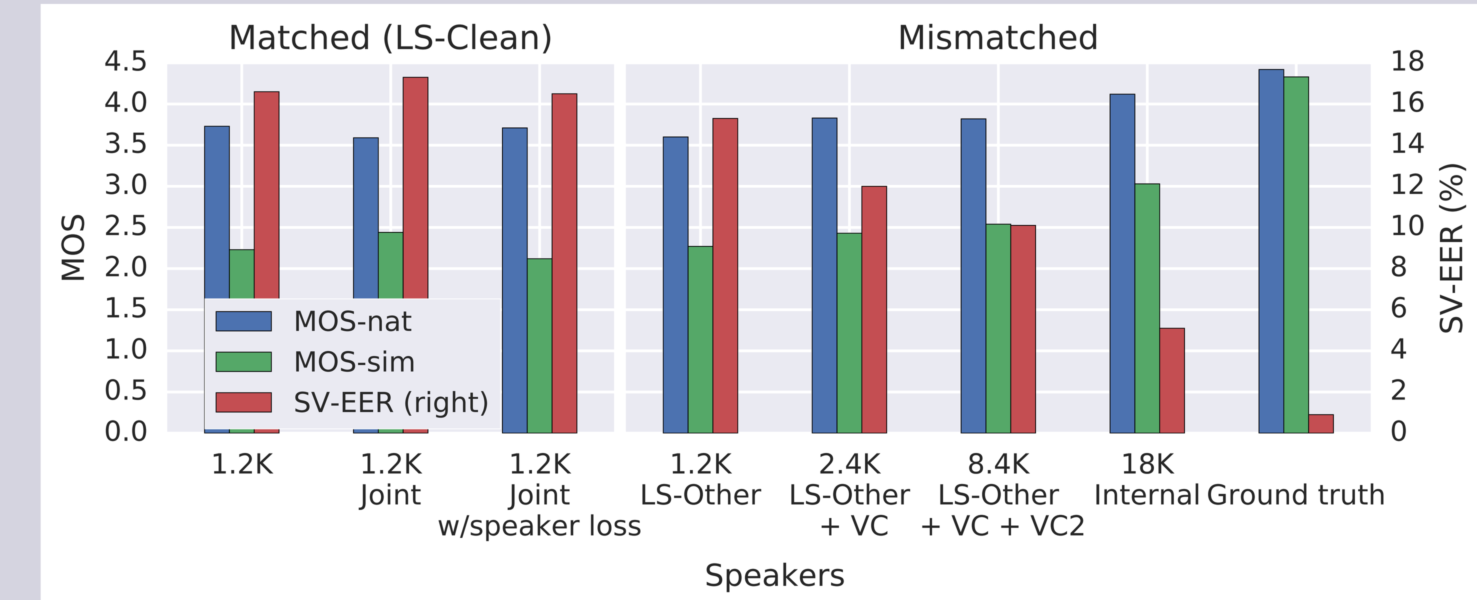


- Real and synthetic utterances from the same speaker (same color) are consistently close
- But real and synthetic utterances consistently form distinct clusters (right)
- SV-EER of 2.9% after enrolling 10 real LibriSpeech speakers and 10 synthetic versions
  - i.e. synthetic utterances are nearly always closest to other synthetic utterances for the same speaker

⟹ Synthesized speech resembles target speaker, but not well enough to be confusable with real speech

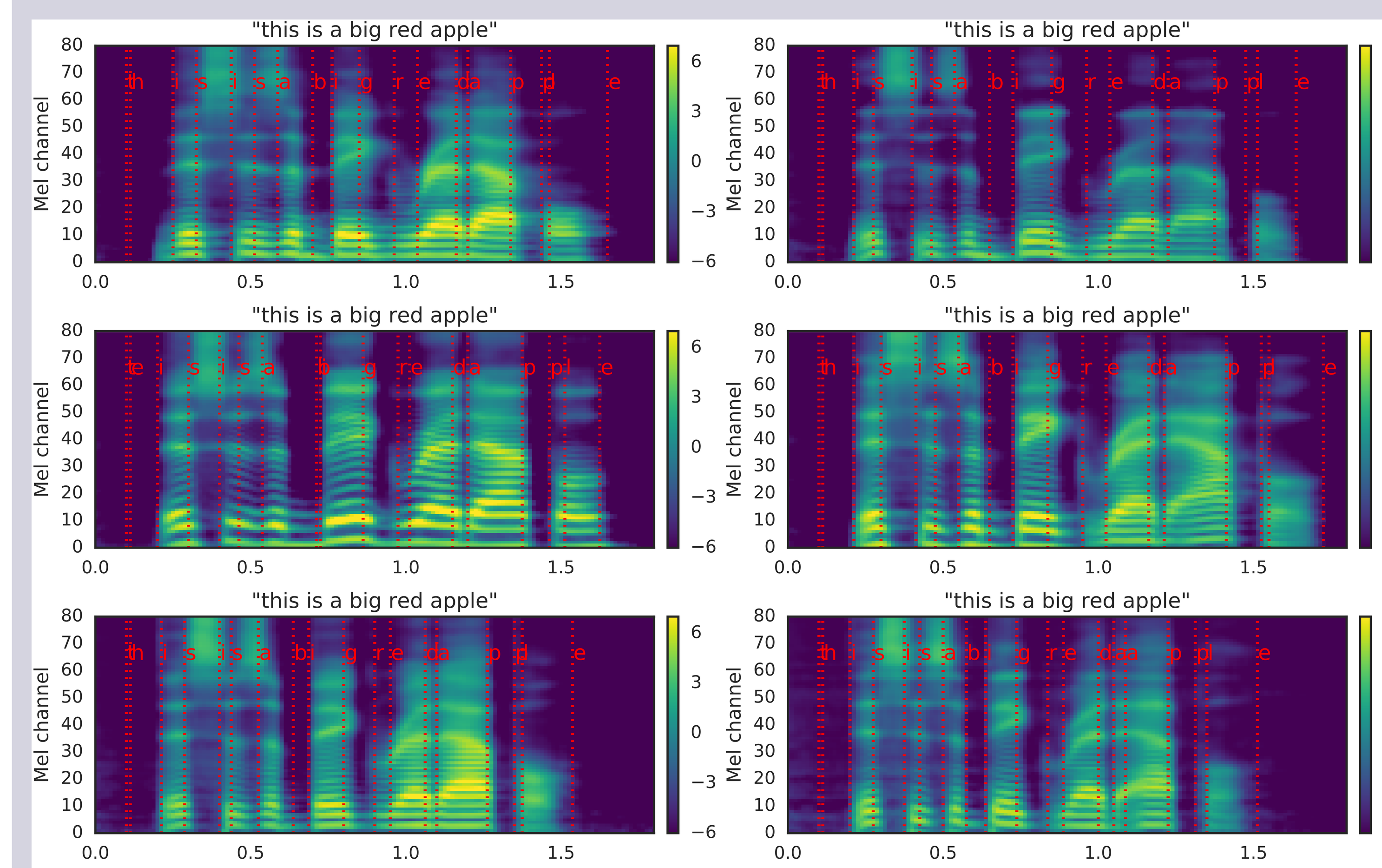
Sound examples at
`https://google.github.io/tacotron/publications/speaker_adaptation`

7. Transfer from speaker encoder



- Compare performance of synthesizer trained on LibriSpeech (LS) conditioned on speaker encoder (SE) trained on different datasets
  - evaluate on previously unseen speakers
- Jointly training SE and synthesizer doesn't improve performance (left)
- Performance improves with number of SE training speakers (right)

8. Fictitious speakers



- Synthesize text conditioned on randomly sampled speaker embeddings
- All samples contain consistent phonetic content, but varied fundamental frequency and speaking rate
- Fictitious speakers speakers are distinct from training speakers

9. References

[1] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou. Neural voice cloning with a few samples. *arXiv preprint arXiv:1802.06006*, 2018.

[2] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf. Fitting new speakers based on a short untranscribed sample. *arXiv preprint arXiv:1802.06984*, 2018.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proc. ICASSP*, 2018.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. *CoRR abs/1609.03499*, 2016.

[5] L. Wan, Q. Wang, A. Papir, and I. L. Moreno. Generalized end-to-end loss for speaker verification. In *Proc. ICASSP*, 2018.