Google

Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

ICASSP 2018, Calgary, April 17, 2018

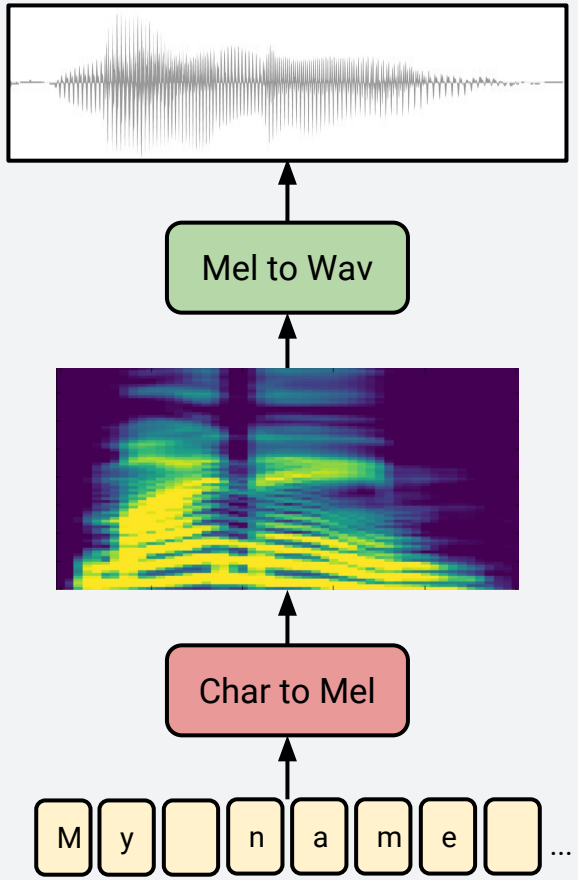
Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu

Tacotron 2

Tacotron 2 is a fully neural text-to-speech system composed of two separate networks.

At the bottom is the feature prediction network, Char to Mel, which predicts mel spectrograms from plain text.

It's followed by a vocoder network, Mel to Wave, that generates waveform samples corresponding to the mel spectrogram features.



Benefits of Tacotron 2

Easy to Get Started With

Tacotron 2 makes it easy to get started with TTS. There is no need for labelled phoneme, duration, or pitch data. **Tacotron 2 can be trained with just the audio and text transcript.**

Decoupling of Content and Audio Quality

The mel spectrogram captures all content information, such as pronunciation, prosody, and speaker identity. Changes to the Char to Mel network only affects content, and changes to the Mel to Wave network only affects audio quality.

Rapid / Parallel Development

The two networks can be improved independently or in parallel. It is not necessary to train both networks to evaluate small changes to one of them.

Caveats

Pronunciation

Tacotron 2 learns pronunciation from the training data. While it can extrapolate quite well to unseen words, it will make mistakes on words with irregular pronunciation.

Textnorm

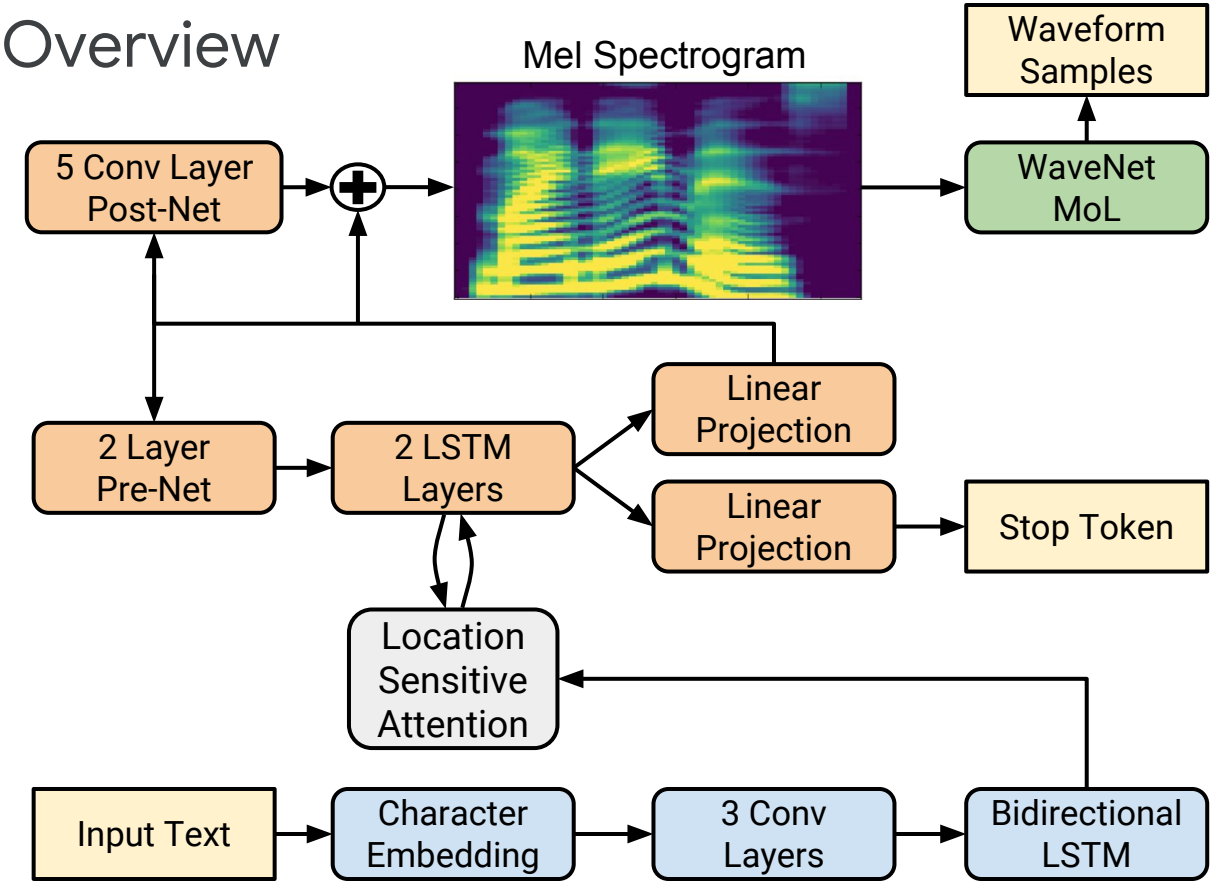
Tacotron 2 has only been trained on verbalized text. I.e., Currency, dates, phone numbers, etc. are written out the way they are spoken. It's unclear how Tacotron 2 would do on the full end-to-end TTS task.

Tweaking Output

It is difficult to adjust the speed or pitch of a mel spectrogram, or to modify the duration of individual phonemes.

Setup

Network Overview



Training

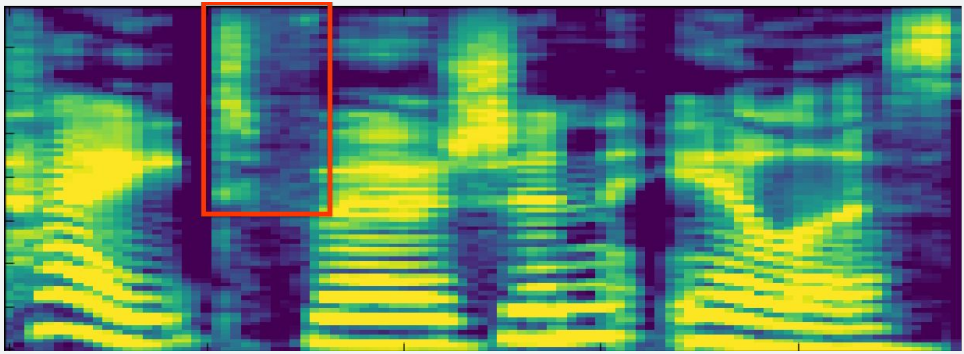
- Char to Mel and Mel to Wave networks trained separately, with independent hyperparameters.
- Teacher-forcing for training.
- Char to Mel: L2 loss on predicted vs groundtruth mel spectrograms.
- Mel to Wave: mixture of logistics loss[1,2].

[1] Salimans, Tim, et al. "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications."
[2] Oord, Aaron van den, et al. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis.", Section 2.1

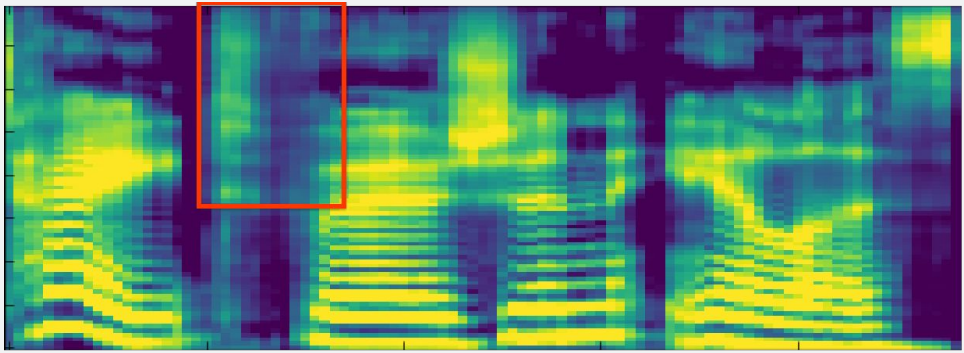
Mel Spectrogram

- tf.contrib.signal.stft()
  tf.contrib.signal.linear_to_mel_weight_matrix()
- **L2 loss drives predictions towards the mean**, which results in **oversmoothed** spectrograms. Mel to Wave network trained on groundtruth spectrograms does not handle this well!
- Solution: train Mel to Wave on predicted spectrograms generated in teacher-forcing mode.

Groundtruth

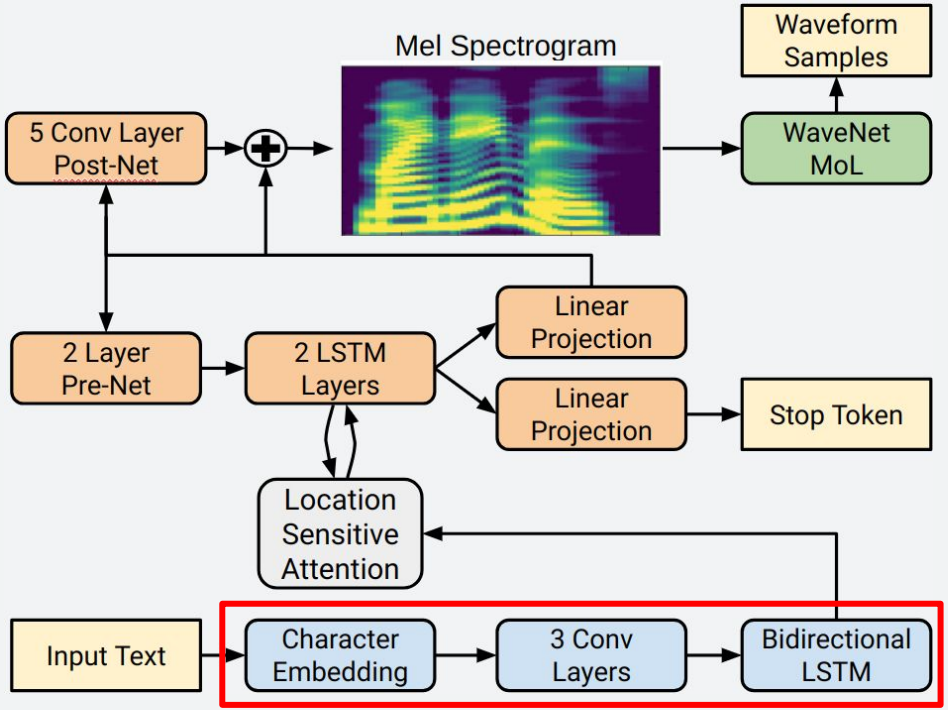


Predicted



Char to Mel - Encoder

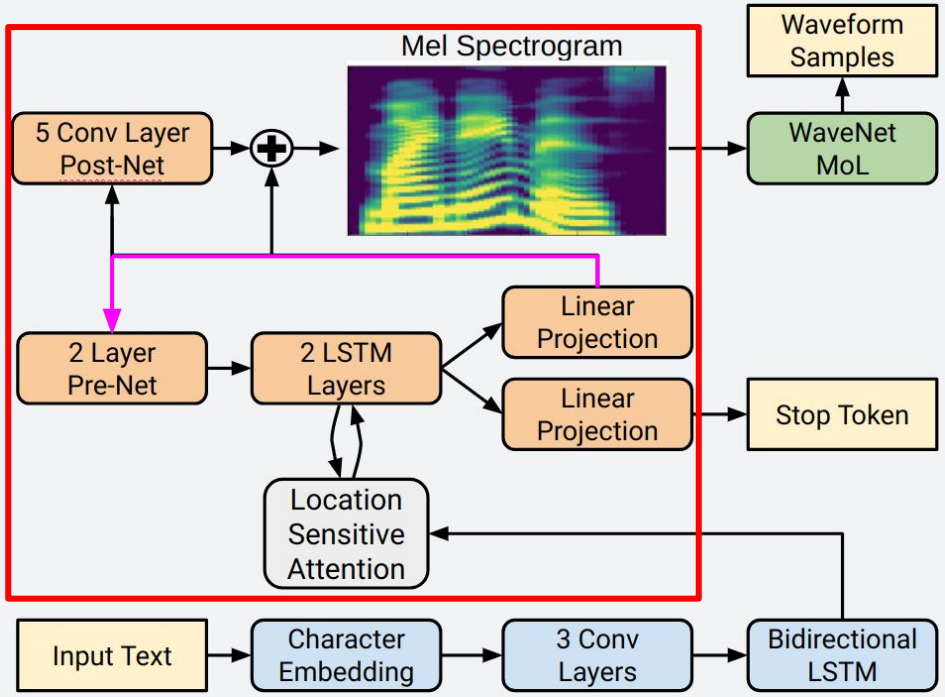
Standard encoder architecture.



Char to Mel - Decoder

Location Sensitive Attention[3].

At each timestep: predict Stop Token in [0, 1].
If >0.5, halt generation. Binary cross-entropy
loss with target = 1 only on the last frame.

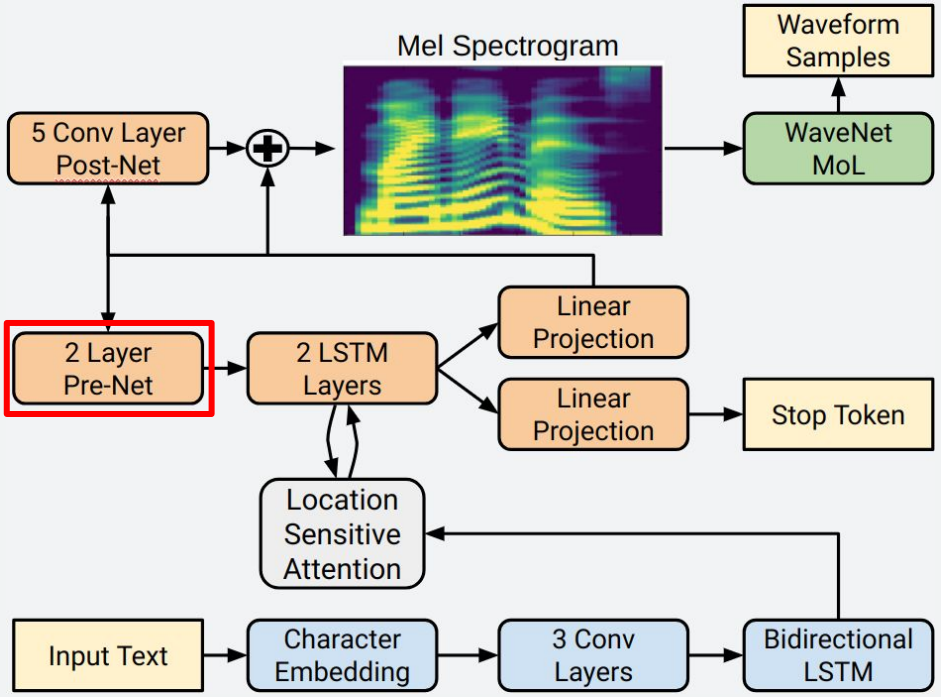


[3] Chorowski, Jan K., et al. "Attention-based models for speech recognition."

Char to Mel - Pre Net

2 fully-connected layers acting as information bottleneck forces decoder to attend to encoder outputs.

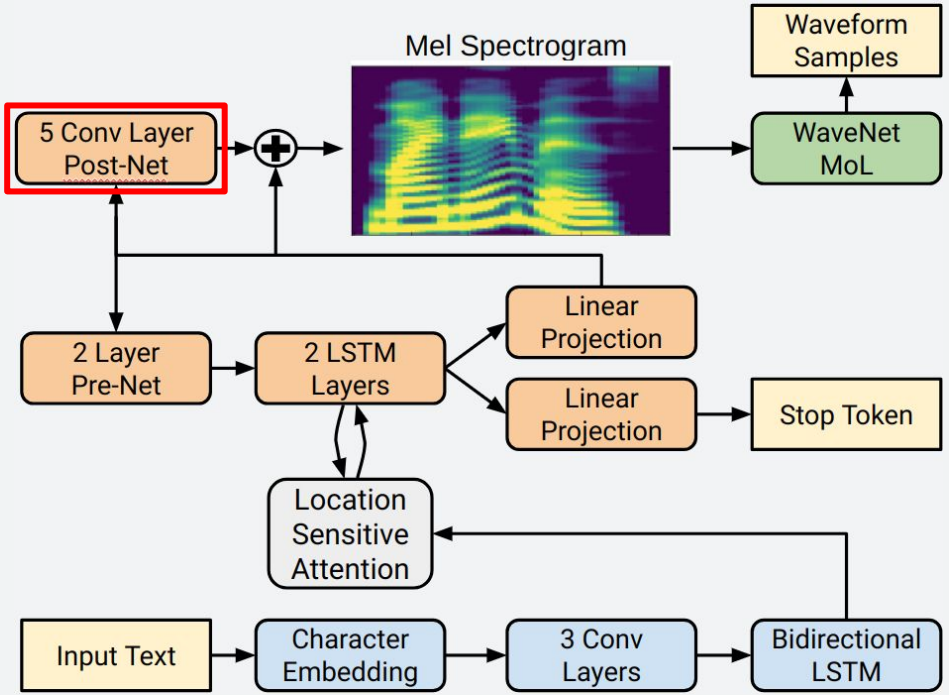
Dropout during inference to induce variation in outputs.



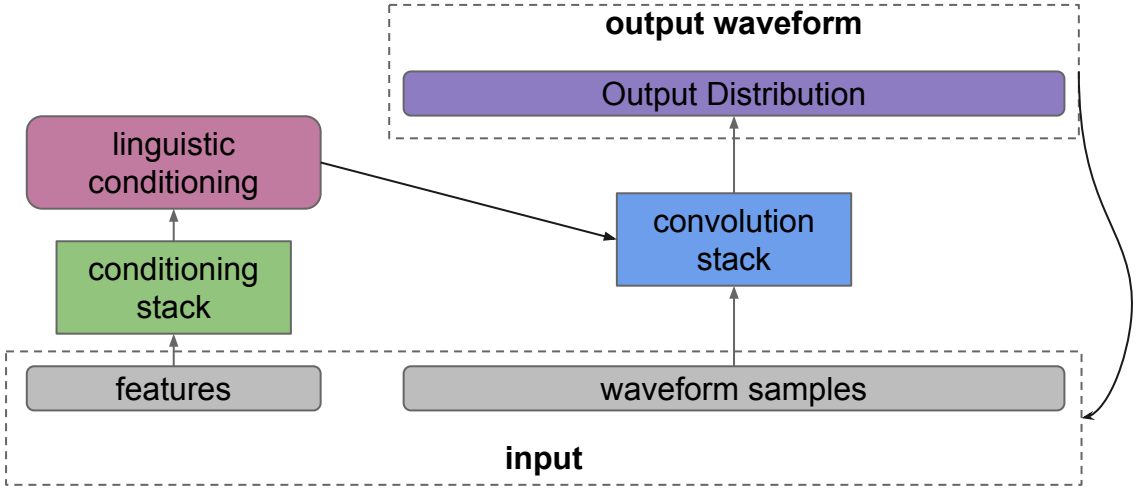
Char to Mel - Post Net

Adds a residual to the spectrogram after all the spectrogram frames are predicted.

Final loss = L2(before postnet) + L2(after postnet) + BCE(stop token)

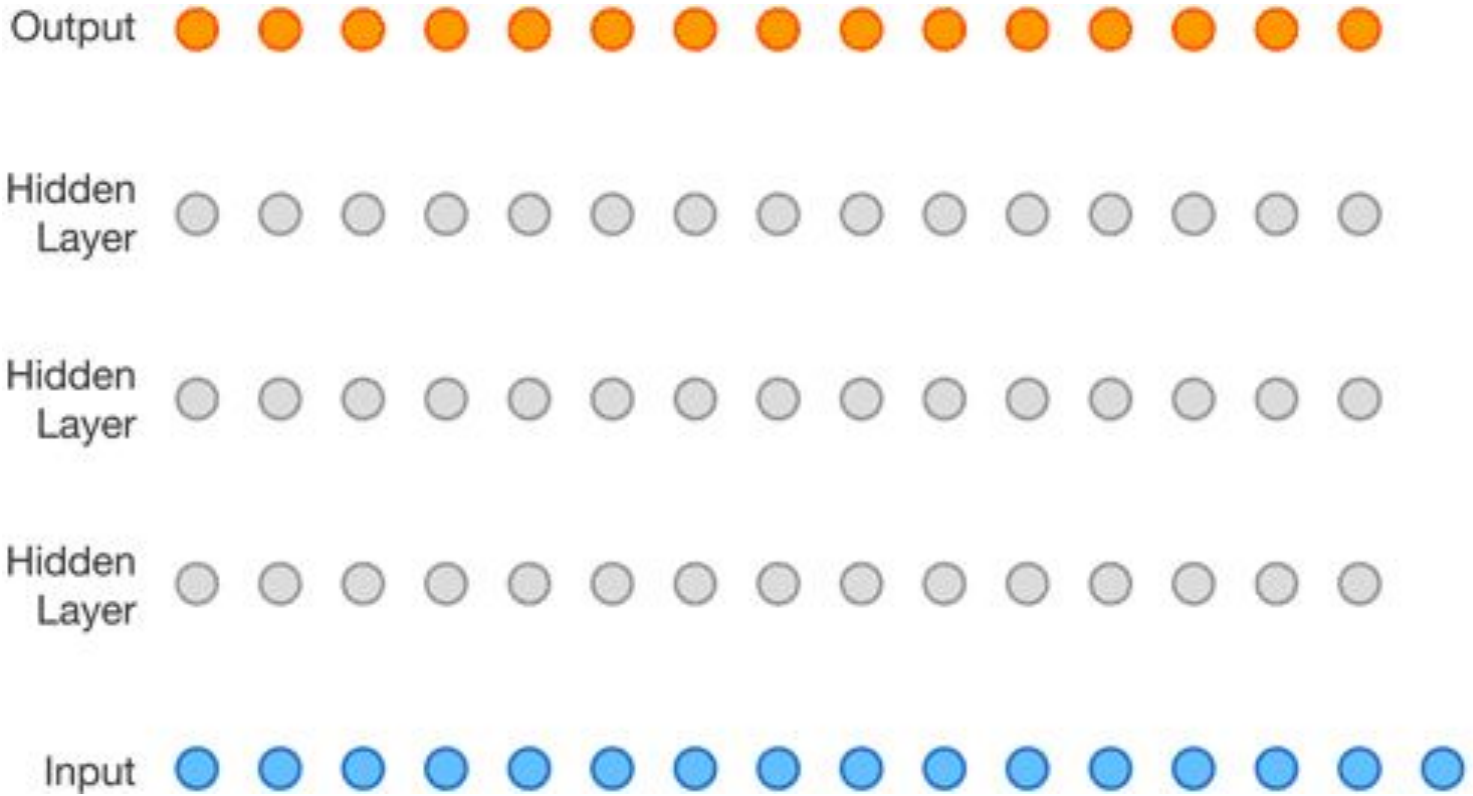


Mel to Wave (WaveNet[4])



[4] Van Den Oord, Aaron, et al. "WaveNet: A generative model for raw audio."

Mel to Wave (WaveNet[4])



J. Shen, et al. | Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

Google

Results

Naturalness Evaluation

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems.

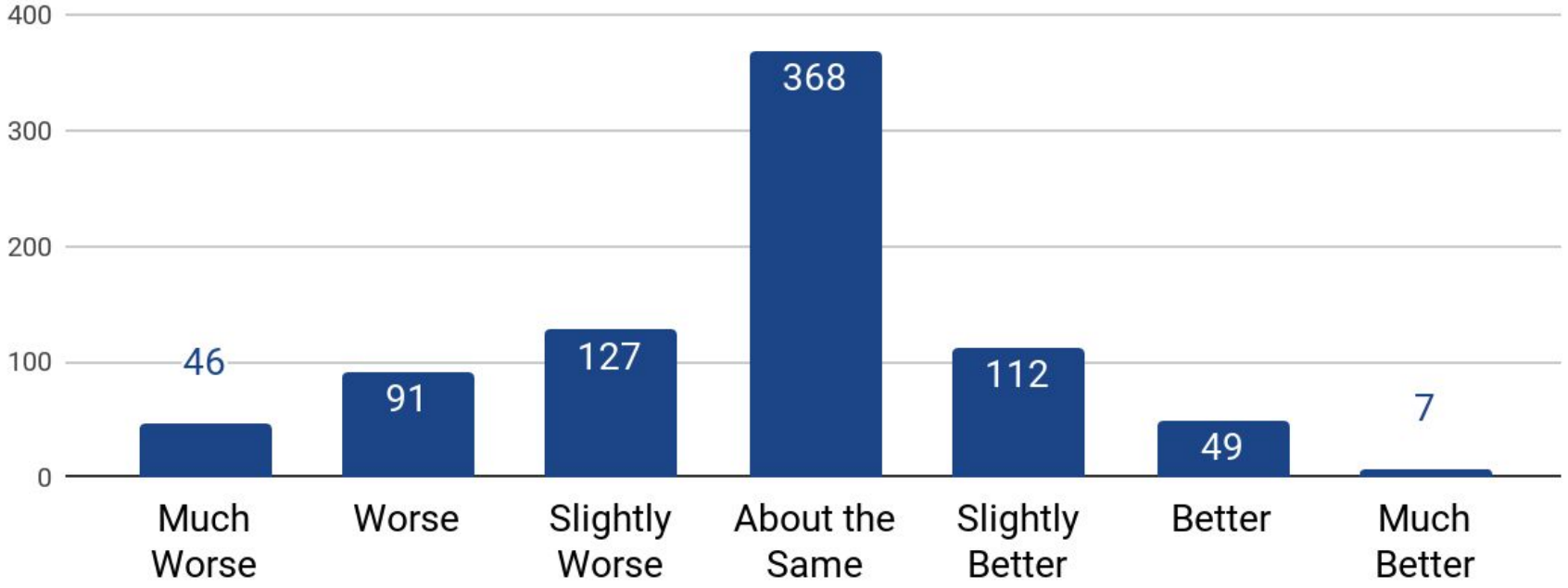
Reducing Size of Mel to Wave Network

The mel spectrogram contains all content information, so the Mel to Wave network doesn't need to do as much work.

Total layers	Num cycles	Dilation cycle size	Receptive field (samples / ms)	MOS
30	3	10	6,139 / 255.8	4.526 ± 0.066
24	4	6	505 / 21.0	4.547 ± 0.056
12	2	6	253 / 10.5	4.481 ± 0.059
30	30	1	61 / 2.5	3.930 ± 0.076

Table 4. WaveNet with various layer and receptive field sizes.

Is Tacotron 2 More Natural Than Recorded Speech



Thank You

More samples can be found at
https://google.github.io/tacotron/publications/tacotron2
or Google "Tacotron 2 samples"

Google

Additional Slides

Training Data

We trained on an internal US English dataset which contains 24.6 hours of professionally recorded speech from a single professional female speaker.

The data extremely high quality (same recording conditions and volume levels, anechoic chamber, no sources of noise) and has consistent and realistic prosody.

Tacotron vs Tacotron 2