

Neural Text-to-Speech

Russel Shawn Dsouza



Electronics and Communications Engg.
National Institute of Technology Karnataka
Surathkal, India - 575025

October 9, 2019

Speech synthesis

Artificial production of human speech

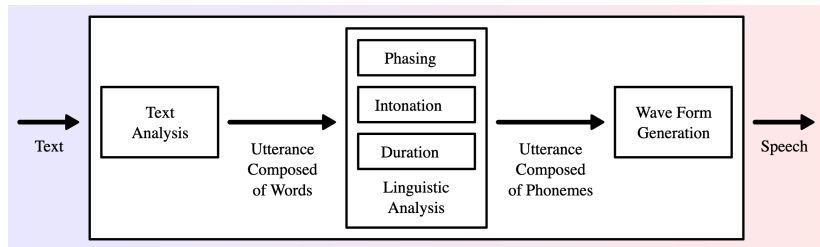


Figure: A typical text-to-speech system¹

¹Andy0101, *A typical text-to-speech system*, https://commons.wikimedia.org/wiki/File:TTS_System.svg, [Online; accessed 10/08/2019], 2010.

History of speech synthesis

Concatenative

- ▶ Extract samples from large database of human speech

Parametric

- ▶ Simulate human voice using a parametric function

Neural

- ▶ Artificially generate human voice using neural networks

Approaches in Neural text-to-speech

- ▶ LSTM
- ▶ WaveNet
- ▶ WaveNet based

WaveNet

A deep neural network for generating raw audiowaveforms.

- ▶ Probabilistic
- ▶ Autoregressive
- ▶ Beats all previously known methods



Figure: Time domain representation of 1 second of generated speech

WaveNet: Architecture

- ▶ Dilated convolution
- ▶ μ law companding
- ▶ Gated activation
- ▶ Residual and skip connection
- ▶ Conditional wavenets

1. Dilated Convolution

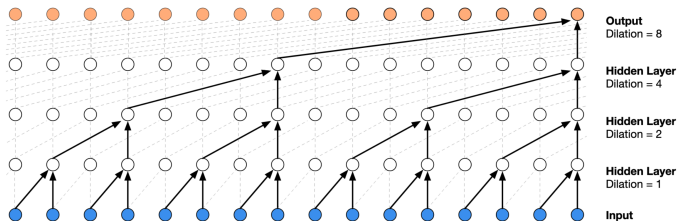


Figure: Stack of dilated causal convolution layers²

²A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, "WaveNet: A Generative Model for Raw Audio," *en, arXiv:1609.03499 [cs]*, Sep. 2016, arXiv: 1609.03499. [Online]. Available: <http://arxiv.org/abs/1609.03499> (visited on 10/08/2019).

2. μ -law companding

$$f(x_t) = \text{sign}(x_t) \frac{\ln(1 + \mu|x_t|)}{\ln(1 + \mu)}$$

where, x_t is the time domain speech signal

3. Gated activation

$$\mathbf{z} = \tanh(W_{f,k} * \mathbf{x}) \circledast \sigma(W_{g,k} * \mathbf{x})$$

4. Residual and skip connections

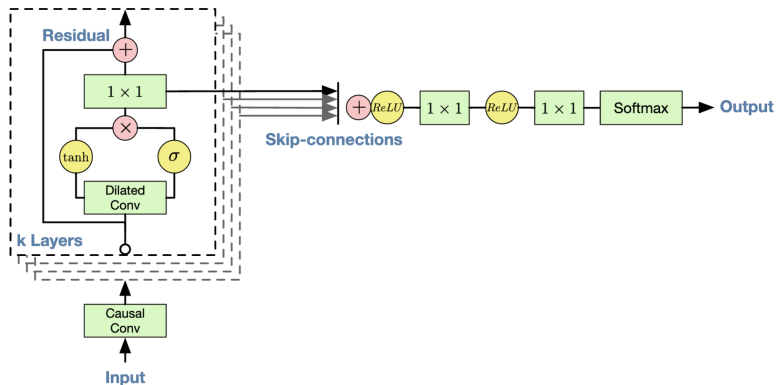


Figure: Overview of residual block and entire architecture³

³A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, "WaveNet: A Generative Model for Raw Audio," *en, arXiv:1609.03499 [cs]*, Sep. 2016, arXiv: 1609.03499. [Online]. Available: <http://arxiv.org/abs/1609.03499> (visited on 10/08/2019).

5. Conditional WaveNets

$$p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^T p(x_t|x_1, \dots, x_{t-1}, \mathbf{h})$$

WaveNet: Pros and Cons

Pros

- ▶ Fast training

Cons

- ▶ Slow inference

Tacotron 2: Architecture

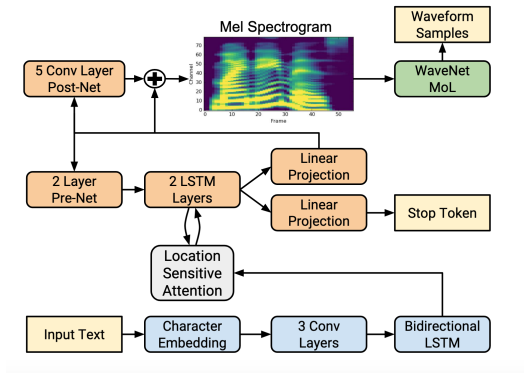


Figure: Block diagram of Tacotron 2 system architecture⁴

⁴J. Shen, R. Pang, R. J. Weiss, et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," en, *arXiv:1712.05884 [cs]*, Dec. 2017, arXiv: 1712.05884. [Online]. Available: <http://arxiv.org/abs/1712.05884> (visited on 10/08/2019).

Mel spectrogram

- ▶ Related to the short-time Fourier transform (STFT) magnitude
- ▶ Obtained by applying a nonlinear transform to the frequency axis of the STFT
- ▶ Emphasizes details in lower frequencies
- ▶ De-emphasizes high frequency details

Features derived from the mel scale have been used as an underlying representation for speech recognition for many decades.⁵

⁵S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

Tacotron 2: Training

Feature detection

- ▶ Maximum likelihood training procedure
- ▶ Batch size = 64 on a single GPU
- ▶ Adam optimizer w/
 $\beta_1 = 0.9$, $\beta_2 = 0.999$,
 $\epsilon = 10^{-6}$
- ▶ LR = 10^{-3} , exponentially
decaying to 10^{-5}
- ▶ Warmup training till 50,000
iterations
- ▶ L2 regularization with
weight 10^{-6}

WaveNet

- ▶ Batch size = 128 on 32
GPUs
- ▶ Adam optimizer w/
 $\beta_1 = 0.9$, $\beta_2 = 0.999$,
 $\epsilon = 10^{-8}$
- ▶ Fixed LR = 10^{-4}
- ▶ Exponentially-weighted
moving average of the
network parameters over
update steps with a decay of
0.9999
- ▶ Scaling by 127.5
- ▶ US English dataset

Tactron 2: Evaluation

Tacotron 2: Reported results

Tacotron 2: Improvements over WaveNet

Summary