Tacotron 2

Natural TTS synthesis by conditioning WaveNet on Mel Spectrogram predictions

Russel Shawn Dsouza (171EC143)

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

October 9, 2019

Overview

Speech synthesis

Artificial production of human speech

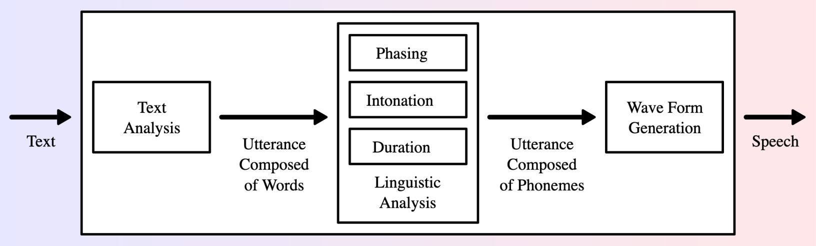

Figure: A typical text-to-speech system[1]

[1] Andy0101, *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[2]

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

[2] V. Delić, Z. Perić, M. Sečujski, *et al.*, "Speech Technology Progress Based on New Machine Learning Paradigm," en, *Computational Intelligence and Neuroscience*, vol. 2019, pp. 1–19, Jun. 2019, ISSN: 1687-5265, 1687-5273. DOI: 10.1155/2019/4368036. [Online]. Available: https://www.hindawi.com/journals/cin/2019/4368036/ (visited on 10/08/2019).

WaveNet

A deep neural network for generating raw audio waveforms.

- Probabilistic
- Autoregressive
- Beats all previously known methods

Figure: Time domain representation of 1 second of generated speech[3]

[3] D. Blog, *WaveNet: A generative model for raw audio*, https://deepmind.com/blog/article/wavenet-generative-model-raw-audio, [Online; accessed 10/08/2019], 2016.

WaveNet: Architecture

Important Components

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

1. Dilated Causal Convolution

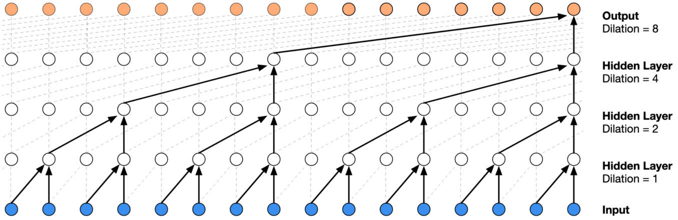

Figure: Stack of dilated causal convolution layers[4]

[4] A. 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[5]

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

$-1 < x_t < 1$ is the time domain speech signal,
$\mu = 255$

[5] Cisco, *Waveform coding techniques*, https://www.cisco.com/c/en/us/support/docs/voice/h323/8123-waveform-coding.html, [Online; accessed 10/09/2019], 2008.

3. Gated activation[6]

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

$*$ $\rightarrow$ convolution,
$\circledast$ $\rightarrow$ element-wise multiplication,
$\sigma(.)$ $\rightarrow$ sigmoid function,
k $\rightarrow$ layer index,
f $\rightarrow$ filter,
g $\rightarrow$ gate,
W $\rightarrow$ learnable convolution filter

[6] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, *et al.*, "Conditional Image Generation with PixelCNN Decoders," *arXiv:1606.05328 [cs]*, Jun. 2016, arXiv: 1606.05328. [Online]. Available: `http://arxiv.org/abs/1606.05328` (visited on 10/08/2019).

4. Residual and skip connections

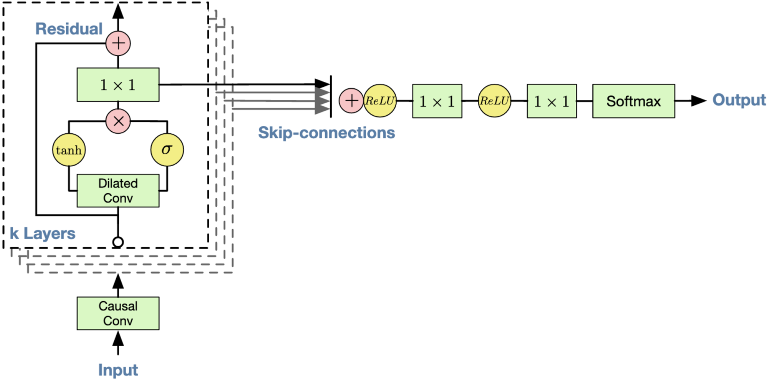

Figure: Overview of residual block and entire architecture[7]

[7] A. 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

Given an additional input $\mathbf{h}$, WaveNets can model the conditional distribution $p(\mathbf{x}|\mathbf{h})$ of the audio given the input,

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

WaveNet: Reported results

Reported results

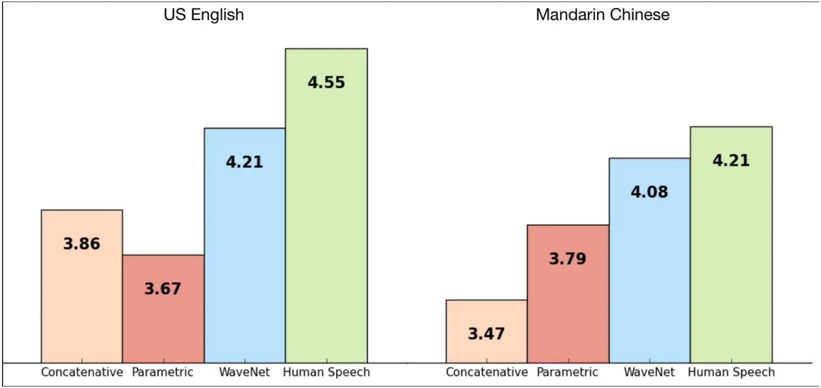

Figure: Mean Opinion Scores (MOS) for English and Mandarin[8]

[8] D. Blog, *WaveNet: A generative model for raw audio*, https://deepmind.com/blog/article/wavenet-generative-model-raw-audio, [Online; accessed 10/08/2019], 2016.

Tacotron 2: Architecture

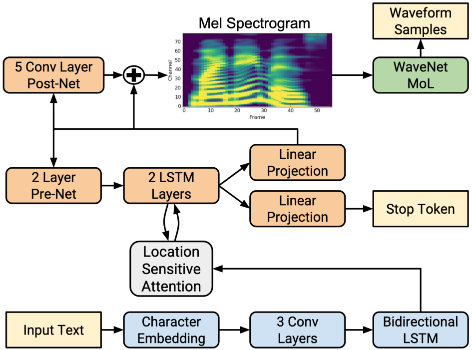

Figure: Block diagram of Tacotron 2 system architecture[9]

[9] J. 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)
- 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.[10]

[10] S. 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 network

- 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

Tacotron 2: Evaluation

- 100 random examples from test set sent to Mechanical Turk
- Each sample is rated by atleast 8 raters
- Scores on a scale of 1 to 5 with 0.5 increments

Tacotron 2: Reported results

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)	$\mathbf{4.526 \pm 0.066}$

Figure: Mean Opinion Scores (MOS)[11]

[11] J. 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).

Conclusions and future strategies

- More general models
- More languages
- Names, abbreviations, context require more work
- Better evaluation and testing required

References I

1. A. 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. J. 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)
3. K. KIM, *WaveNet: Increasing reception field using dilated convolution*, https://medium.com/@kion.kim/wavenet-a-network-good-to-know-7caaae735435, [Online; accessed 10/09/2019]

References II

4. S. Kumar, *Understanding WaveNet architecture*, https://medium.com/@satyam.kumar.iiitv/understanding-wavenet-architecture-361cc4c2d623, [Online; accessed 10/09/2019]
5. J. Singh, *WaveNet: Google Assistant's Voice Synthesizer*, https://towardsdatascience.com/wavenet-google-assistants-voice-synthesizer-a168e9af13b1, [Online; accessed 10/09/2019]
6. Q. Yongliang, *Behind WaveNet*, https://ctmakro.github.io/site/on_learning/audio/wavenet_arch.html, [Online; accessed 10/09/2019]
7. D. Mwiti, *A 2019 guide to speech synthesis with deep learning*, https://heartbeat.fritz.ai/a-2019-guide-to-speech-synthesis-with-deep-learning-630afcafb9dd, [Online; accessed 10/09/2019]