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

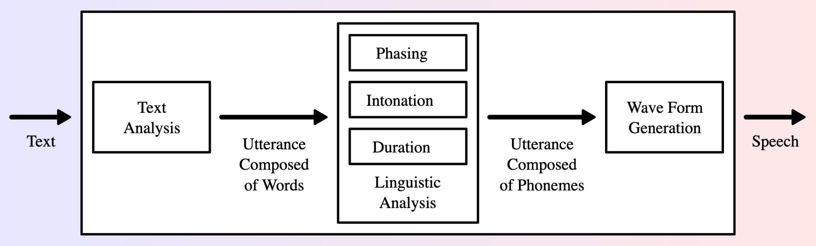

Figure: A typical text-to-speech system[1]

[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

Concatenative

- Large database of human speech used

Parametric

- Simulate human voice using a function

Neural

- 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
- Context stacks

1. Dilated Convolution

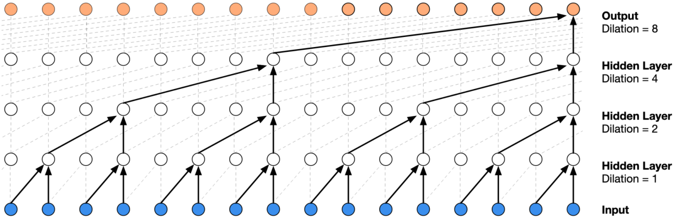

Figure: Stack of dilated causal convolution layers[2]

[2] 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

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

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

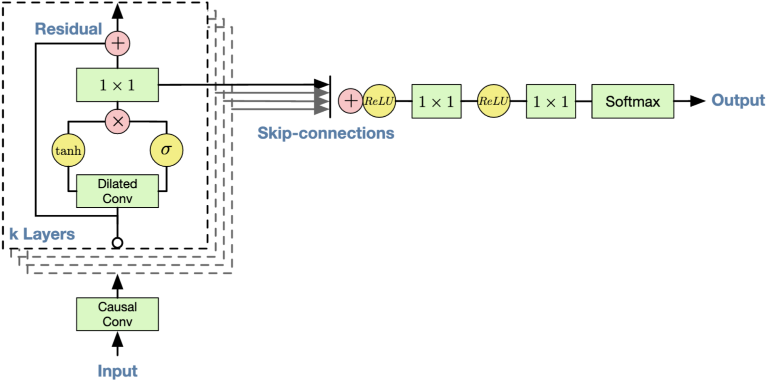

Figure: Overview of residual block and entire architecture[3]

[3] 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

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

6. Context Stacks

WaveNet: Pros and Cons

Pros

- Fast training

Cons

- Slow inference

Tacotron 2

Tacotron 2: Architecture

Tacotron 2: Training

Tacotron 2: Reported results

Tacotron 2: Improvements over WaveNet

Neural TTS: The future

Summary

Conclusion