WaveNet: A Generative Model for Raw Audio

*Aäron van den Oord, Sander Dieleman, Heiga Zen[†], Karen Simonyan, Oriol Vinyals,
Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu*

DeepMind, London, United Kingdom [†] Google, London, United Kingdom

1. Abstract

This demo presents WaveNet [1], a deep generative model of raw audio waveforms. We show that WaveNets are able to generate speech which mimics any human voice and which sounds more natural than the best existing Text-to-Speech (TTS) systems, reducing the gap in subjective quality relative to natural speech by over 50%. We also demonstrate that the same network can be used to synthesize other audio signals such as music, and present some striking samples of automatically generated piano pieces. WaveNets open up a lot of possibilities for text-to-speech, music generation and audio modelling in general.

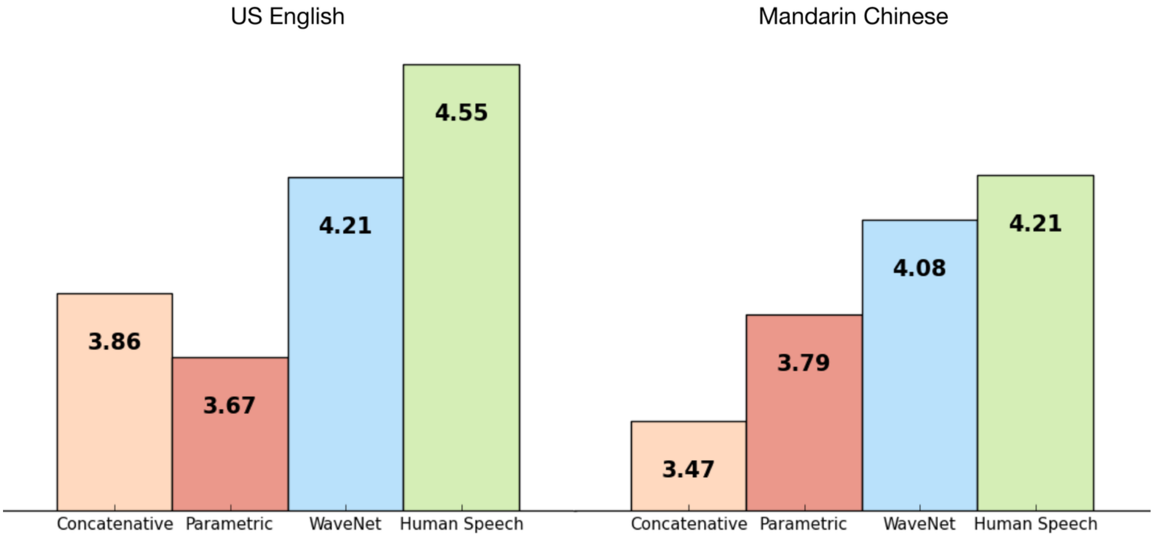

Figure 1: *Subjective 5-scale Mean Opinion Scores (MOS) of speech samples from HMM-driven unit selection concatenative [2], LSTM-RNN parametric [3], and WaveNet [1] -based speech synthesizers, and human natural speech. WaveNet improved the previous state of the art significantly, reducing the gap relative to natural speech by over 50% in both languages.*

2. References

[1] 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," `https://drive.google.com/file/d/0B3cxcnOkPx9AeWpLVXhkTDJINDQ/view`, 2016.

[2] X. Gonzalvo, S. Tazari, C.-A. Chan, M. Becker, A. Gutkin, and H. Silen, "Recent advances in Google real-time HMM-driven unit selection synthesizer," in *Proc. Interspeech*, 2016, pp. 2238–2242.

[3] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," in *Interspeech*, 2016, pp. 2273–2277.