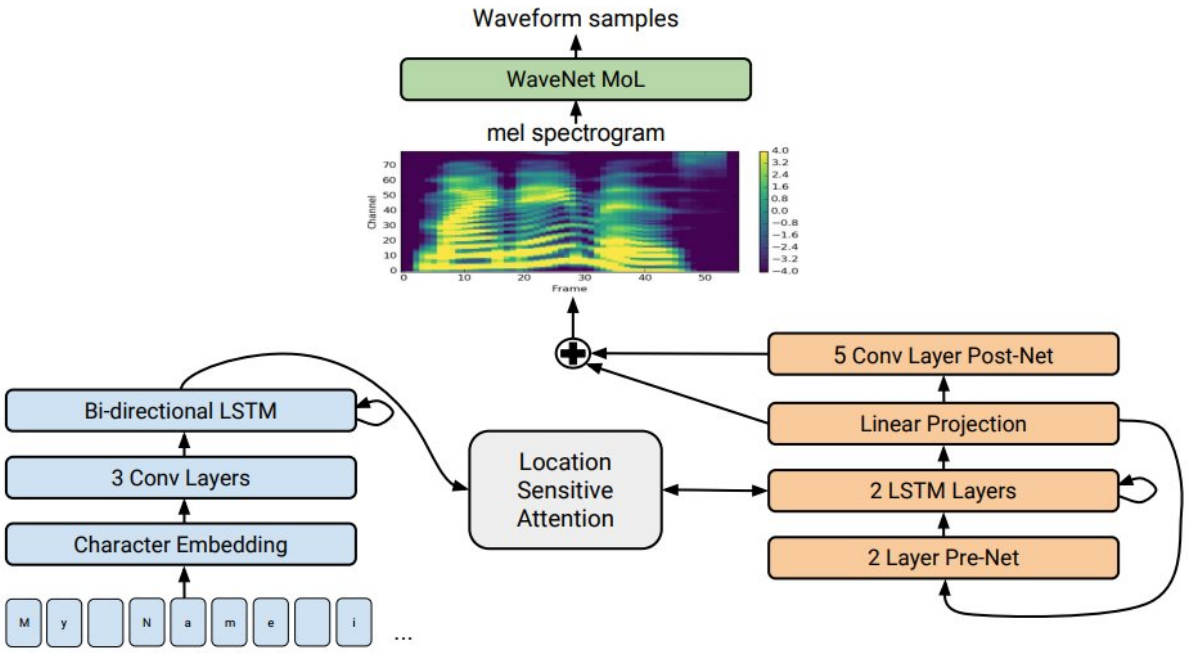
Google offers update on its human-like text-to-speech system

December 29 2017, by Bob Yirka



A detailed look at Tacotron 2's model architecture. The lower half of the image describes the sequence-to-sequence model that maps a sequence of letters to a spectrogram. For technical details, please refer to the paper. Credit: Google

Google has offered interested tech enthusiasts an update on its Tacotron text-to-speech system via blog post this week. In the post, the team describes how the system works and offers some audio samples, which Ruoming Pang and Jonathan Shen, authors of the post, claim were

comparable to professional recordings as judged by a group of human listeners. The authors have also written a paper with the rest of their Google teammates describing their efforts, and have posted it to the *arXiv* preprint server.

For many years, scientists have been working to make computer generated speech sound more human and less robotic. One part of that mission is developing text-to-speech (TTS) applications, as the authors note. Most people have heard the results of TTS systems, such as the automated voice systems used by many corporations to field customer calls. In this new effort, the group at Google has combined what it learned from its Tacotron and WaveNet projects to create Tacotron 2—a system that takes the science to a new level. In listening to the provided samples, it is quite difficult and sometimes impossible to tell if a voice is a human or a TTS system voice.

To achieve this new level of accuracy, the team at Google used a sequence-to-sequence model optimized to work with TTS—it maps arrangements of letters to a series of features that describe the audio. The result is an 80-dimensional spectrogram. That spectrogram is then used as input to a second system that outputs a 24-kHz waveform using an architecture based on WaveNet. Both are neural networks trained using speech examples (from crowdsourcing applications such as Amazon's Mechanical Turk) and their corresponding transcripts. The new system is able to incorporate volume, pronunciation, intonation and speed, allowing for the creation of a much more human-like voice.

The team also notes that they are still working to improve the system, most notably to overcome problems with complex words and to make it work in real time. They would also like to add more emotion to the voice so listeners could actually hear happiness or sadness, for example, or to detect displeasure. Doing so would not only advance the science, but it would make interactions with digital assistants more intimate.

More information: Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, arXiv:1712.05884 [cs.CL] arxiv.org/abs/1712.05884

Abstract

This paper describes Tacotron 2, a neural network architecture for speech synthesis directly from text. The system is composed of a recurrent sequence-to-sequence feature prediction network that maps character embeddings to mel-scale spectrograms, followed by a modified WaveNet model acting as a vocoder to synthesize timedomain waveforms from those spectrograms. Our model achieves a mean opinion score (MOS) of 4.53 comparable to a MOS of 4.58 for professionally recorded speech. To validate our design choices, we present ablation studies of key components of our system and evaluate the impact of using mel spectrograms as the input to WaveNet instead of linguistic, duration, and F0 features. We further demonstrate that using a compact acoustic intermediate representation enables significant simplification of the WaveNet architecture.

Citation: Google offers update on its human-like text-to-speech system (2017, December 29) retrieved 16 April 2024 from https://techxplore.com/news/2017-12-google-human-like-text-to-speech.html