Tacotron 2: Natural TTS synthesis by conditioning WaveNet on Mel Spectrogram

Russel Shawn Dsouza (171EC143)

October 9, 2019

Generating speech from text has been a challenging problem for decades despite a lot of research and investigation. Over time, different techniques have dominated the field. Concatenative synthesis and parametric synthesis were the 2 most popular computational speech generation algorithms. However, the audio produced by these systems often sounds muffled and robotic compared to human speech. In 2017, WaveNet, a generative model of time domain waveforms, produced audio quality that rivaled human speech for the first time in history. The inputs to the WaveNet model required significant domain expertise to produce, and involved elaborate text-analysis systems as well as a robust lexicon. Taxotron, a sequence-to-sequence architecture for producing magnitude spectrograms from a sequence of characters, simplifies the traditional speech synthesis pipeline by replacing the production of linguistic and acoustic features with a single neural network trained from data alone. To vocode the resulting magnitude spectrum, Tacotron 1 used the Griffin-Lim algorithm followed by inverse STFT. This simplified the input generation process by produced audio that was inferior to the WaveNet. Tacotron 2, introduced in 2018, is an entirely neural approach to speech synthesis where the vocoder of Tacotron 1 is replaced by a WaveNet vocoder. This network beats the state of the art WaveNet network and produces subjectively more natural sounding human voice as rated on Amazon Mechanical Turk.