Tacotron 2 is a unified neural approach that uses a network similar to Tacotron to convert text into mel-spectograms and then uses a WaveNet vocoder to convert the spectrograms into speech that sounds natural. A 512-dimensional embedding is used as input, which is then processed by a convolutional network that encodes long-term context. The encoded features are created by feeding the final layer's output into a bidirectional LSTM. The encoder output is subsequently passed to the "location sensitive attention" module, which integrates the attention weights collected during previous decoder time steps. Based on the encoded input, two long short-term memory (LSTM) units predict a mel-spectogram, and this unit acts as the decoder. The time domain of the final spectrogram is mapped using LSTM, which is made feasible by the network that creates spectra's usage of the short-time Fourier transform. Tacotron 2 creates waveform audio by translating freeform text to mel-scale spectrograms and then sending the data through a modified WaveNet-based vocoder model. The WaveNet vocoder is a modification of the original WaveNet concept that generates audio using spectrograms rather than text. They use 30 dilated convolution layers, which are divided into three dilatation cycles and two upsampling layers.

Natural voice production from text is still a challenging task. The purpose of this research is to develop a technique for generating human-like voice from written text. However, in prior attempts, the audio created by such systems sounded muffled and unnatural when compared to human speech. This approach allows for thorough TTS learning from character sequences and speech waveforms, producing speech that is highly close to human voice. The system is composed of two major components: an

attention-based recurrent sequence-to-sequence feature prediction network that predicts a collection of mel-spectrogram frames from an input character sequence and a modified WaveNet that generates time-domain waveform samples from the predicted mel-spectrogram frames. The authors explain why mel-spectograms perform effectively in TTS systems. We may train the two components individually by utilizing a representation that can be easily produced from the time-domain waveform. Because this representation is phase invariant within each frame, it is smoother than waveform samples and can be learned more easily using the mean squared error loss. Despite not providing any additional features to feed the vocoder, as did Deep Voice, the quality of the retrieved samples was far greater than that created using WaveNet. They delivered considerably superior extracted samples than Deep Voice, yet they did not add any new features to the vocoder. The training dataset is an internal dataset of female-spoken US English. The real-world target is unknown while creating speech inferentially. As a result, rather than the teacher-enforced configuration used for training, the projected output of the preceding phase is input during the decoding process. The authors demonstrated that using mel-spectograms rather than linear-spectograms or linguistic characteristics was critical to improving the model's performance. The authors also examined the types of errors that the model produces in the test set and discovered that the most common error is unusual pitch or emphasis and suggested that this should be the primary area of development to work on. They also tested the model on news headlines to see how it generalized and discovered that it sometimes encountered pronunciation issues. Speech is synthesized to Tacotron-level prosody and WaveNet-level audio quality by the system. The system can be trained directly from

data without the need for extensive feature engineering, and it provides state-of-the-art sound quality that is comparable to real human speech. Speech synthesized by Tacotron 2 is up to Tacotron-level prosody and WaveNet-level audio quality. The system can be trained directly from data without the need for extensive feature engineering, and it provides state-of-the-art sound quality that is comparable to real human speech. This paper broadens the scope of TTS study. It would be fascinating in future work if the prosody and emotional tone of the voice could be incorporated into a model.