WaveNet and other synthetic voices

Text-to-Speech creates raw audio data of natural, human speech. That is, it creates audio that sounds like a person talking. When you send a synthesis request to Text-to-Speech, you must specify a *voice* that 'speaks' the words.

There are a wide selection of custom voices available for you to pick from in Text-to-Speech. The voices differ by language, gender, and accent (for some languages). Some languages have multiple voices to choose from. You can see a list of the voices available for speech synthesis in the Text-to-Speech on the Supported Voices (https://cloud.google.com/text-to-speech/docs/voices) page.

The voices offered from Text-to-Speech can also differ in how they are produced, the synthetic speech technology used to create the machine model of the voice. One common speech technology, *parametric text-to-speech*, typically generates audio data by passing outputs through signal processing algorithms known as *vocoders* (https://en.wikipedia.org/wiki/Vocoder). Many of the standard voices available in Text-to-Speech use a variation of this technology.

WaveNet voices

The Cloud Text-to-Speech API also offers a group of premium voices generated using a *WaveNet model*, the same technology used to produce speech for Google Assistant, Google Search, and Google Translate. WaveNet technology provides more than just a series of synthetic voices: it represents a new way of creating synthetic speech.

Note: Check the table of supported voice (https://cloud.google.com/text-to-speech/docs/voices) for availability of WaveNet-generated voices in specific languages. The Cloud Text-to-Speech API does not provide access to the voice of the Google Assistant.

A WaveNet generates speech that sounds more natural than other text-to-speech systems. It synthesizes speech with more human-like emphasis and inflection on syllables, phonemes, and words. On average, a WaveNet produces speech audio that people prefer over other text-to-speech technologies.

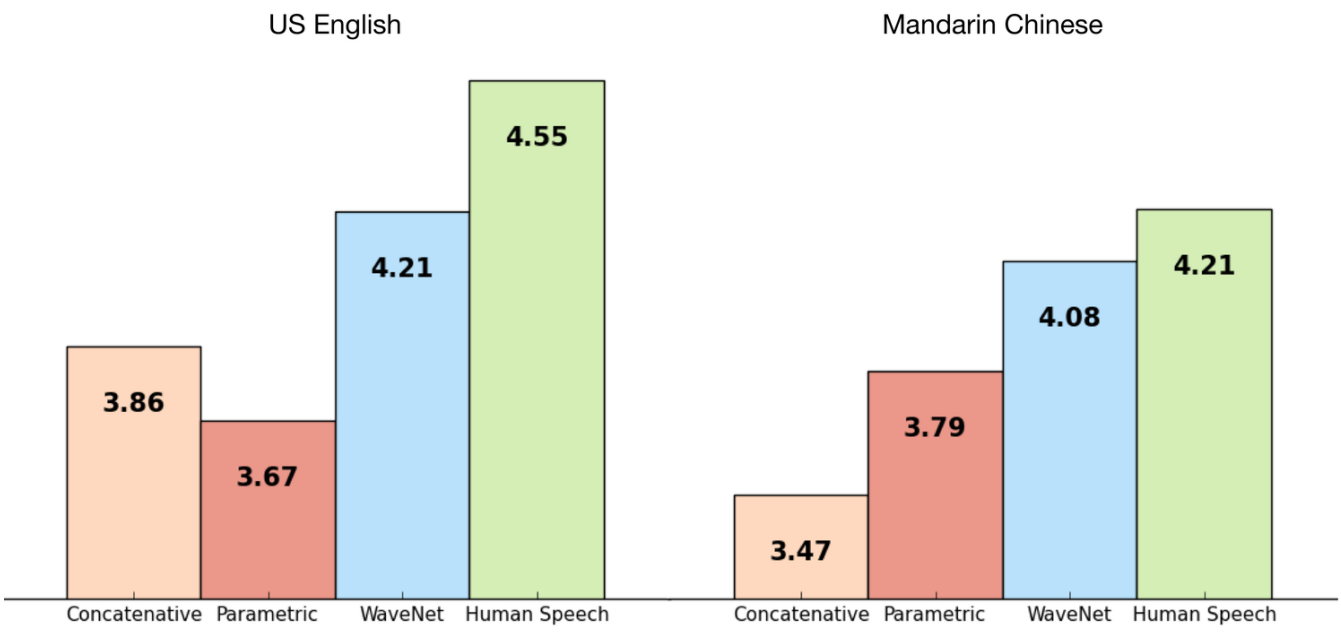

Figure 1. Chart showing comparison of WaveNet to other synthetic voices, human speech.

Unlike most other text-to-speech systems, a WaveNet model creates raw audio waveforms from scratch. The model uses a neural network that has been trained using a large volume of speech samples. During training, the network extracts the underlying structure of the speech, such as which tones follow each other and what a realistic speech waveform looks like. When given a text input, the trained WaveNet model can generate the corresponding speech waveforms from scratch, one sample at a time, with up to 24,000 samples per second and seamless transitions between the individual sounds.

Note: Using WaveNet voices in your text-to-speech synthesis has different pricing than non-WaveNet generated audio. For more details, see the pricing page (https://cloud.google.com/text-to-speech/pricing).

To hear the difference between a Wavenet-generated audio clip and a clip generated by another text-to-speech process, compare the two audio clips below.

0:00 / 0:05

Example 1. High quality, non-WaveNet voice

0:00 / 0:06

Example 2. WaveNet voice

To learn more about WaveNet models, read this blog post by DeepMind (https://deepmind.com/blog/wavenet-generative-model-raw-audio/).

Last updated December 4, 2019.