WaveNet is a deep neural network for generating raw audio. It was created by researchers at London-based artificial intelligence firm DeepMind. The technique, outlined in a paper in September 2016, is able to generate relatively realistic-sounding human-like voices by directly modelling waveforms using a neural network method trained with recordings of real speech. Tests with US English and Mandarin reportedly showed that the system outperforms Google's best existing text-to-speech (TTS) systems, although as of 2016 its text-to-speech synthesis still was less convincing than actual human speech. WaveNet's ability to generate raw waveforms means that it can model any kind of audio, including music.

WaveNet is a type of feedforward neural network known as a deep convolutional neural network (CNN). In WaveNet, the CNN takes a raw signal as an input and synthesises an output one sample at a time. It does so by sampling from a softmax (i.e. categorical) distribution of a signal value that is encoded using μ-law companding transformation and quantized to 256 possible values.

According to the original September 2016 DeepMind research paper *WaveNet: A Generative Model for Raw Audio*, the network was fed real waveforms of speech in English and Mandarin. As these pass through the network, it learns a set of rules to describe how the audio waveform evolves over time. The trained network can then be used to create new speech-like waveforms at 16,000 samples per second. These waveforms include realistic breaths and lip smacks – but do not conform to any language.

WaveNet is able to accurately model different voices, with the accent and tone of the input correlating with the output. For example, if it is trained with German, it produces German speech. The capability also means that if the WaveNet is fed other inputs – such as music – its output will be musical. At the time of its release, DeepMind showed that WaveNet could produce waveforms that sound like classical music.

Content (voice) swapping

According to the June 2018 paper *Disentangled Sequential Autoencoder*,[15] DeepMind has successfully used WaveNet for audio and voice "content swapping": the network can swap the voice on an audio recording for another, pre-existing voice while maintaining the text and other features from the original recording. "We also experiment on audio sequence data. Our disentangled representation allows us to convert speaker identities into each other while conditioning on the content of the speech." (p. 5) "For audio, this allows us to convert a male speaker into a female speaker and vice versa *[...]*." (p. 1) According to the paper, a two-digit minimum amount of hours (c. 50 hours) of pre-existing speech recordings of both source and target voice are required to be fed into WaveNet for the program to learn their individual features before it is able to perform the conversion from one voice to another at a satisfying quality. The authors stress that "*[a]*n advantage of the model is that it separates dynamical from static features *[...]*." (p. 8), i. e. WaveNet is capable of distinguishing between the spoken text and modes of delivery (modulation, speed, pitch, mood, etc.) to maintain during the conversion from one voice to another on the one hand, and the basic features of both source and target voices that it is required to swap on the other.

The January 2019 follow-up paper *Unsupervised speech representation learning using WaveNet autoencoders* details a method to successfully enhance the proper automatic recognition and discrimination between dynamical and static features for "content swapping", notably including swapping voices on existing audio recordings, in order to make it more reliable. Another follow-up paper, *Sample Efficient Adaptive Text-to-Speech*, dated September 2018 (latest revision January 2019), states that DeepMind has successfully reduced the minimum amount of real-life recordings required to sample an existing voice via WaveNet to "merely a few minutes of audio data" while maintaining high-quality results.

Its ability to clone voices has raised ethical concerns about WaveNet's ability to mimic the voices of living and dead persons. According to a 2016 BBC article, companies working on similar voice-cloning technologies (such as Adobe Voco) intend to insert watermarking inaudible to humans to prevent counterfeiting, while maintaining that voice cloning satisfying, for instance, the needs of entertainment-industry purposes would be of a far lower complexity and use different methods than required to fool forensic evidencing methods and electronic ID devices, so that natural voices and voices cloned for entertainment-industry purposes could still be easily told apart by technological analysis.

Applications

At the time of its release, DeepMind said that WaveNet required too much computational processing power to be used in real world applications. As of October 2017, Google announced a 1,000-fold performance improvement along with better voice quality. WaveNet was then used to generate Google Assistant voices for US English and Japanese across all Google platforms. In November 2017, DeepMind researchers released a research paper detailing a proposed method of "generating high-fidelity speech samples at more than 20 times faster than real-time", called "Probability Density Distillation". At the annual I/O developer conference in May 2018, it was announced that new Google Assistant voices were available and made possible by WaveNet; WaveNet greatly reduced the number of audio recordings that were required to create a voice model by modeling the raw audio of the voice actor samples.