WaveNet: A Generative Model for Raw Audio

This post presents **WaveNet**, a deep generative model of raw audio waveforms. We show that WaveNets are able to generate speech which mimics any human voice and which sounds more natural than the best existing Text-to-Speech systems, reducing the gap with human performance by over 50%.

We also demonstrate that the same network can be used to synthesize other audio signals such as music, and present some striking samples of automatically generated piano pieces.

Talking Machines

Allowing people to converse with machines is a long-standing dream of human-computer interaction. The ability of computers to understand natural speech has been revolutionised in the last few years by the application of deep neural networks (e.g., **Google Voice Search**). However, generating speech with computers — a process usually referred to as **speech synthesis** or text-to-speech (TTS) — is still largely based on so-called **concatenative TTS**, where a very large database of short speech fragments are recorded from a single speaker and then recombined to form complete

utterances. This makes it difficult to modify the voice (for example switching to a different speaker, or altering the emphasis or emotion of their speech) without recording a whole new database.

This has led to a great demand for **parametric TTS**, where all the information required to generate the data is stored in the parameters of the model, and the contents and characteristics of the speech can be controlled via the inputs to the model. So far, however, parametric TTS has tended to sound less natural than concatenative, at least for syllabic languages such as English. Existing parametric models typically generate audio signals by passing their outputs through signal processing algorithms known as **vocoders**.

WaveNet changes this paradigm by directly modelling the raw waveform of the audio signal, one sample at a time. As well as yielding more natural-sounding speech, using raw waveforms means that WaveNet can model any kind of audio, including music.

WaveNets

Researchers usually avoid modelling raw audio because it ticks so quickly: typically 16,000 samples per second or more, with important structure at many time-scales. Building a completely autoregressive model, in which the prediction for every one of those samples is influenced by all previous ones (in statistics-speak, each predictive distribution is conditioned on all previous observations), is clearly a challenging task.

However, our **PixelRNN** and **PixelCNN** models, published earlier this year, showed that it was possible to generate complex natural images not only one pixel at a time, but one colour-channel at a time, requiring thousands of predictions per image. This inspired us to adapt our two-dimensional PixelNets to a one-dimensional WaveNet.

The above animation shows how a WaveNet is structured. It is a fully convolutional neural network, where the convolutional layers have various dilation factors that allow its receptive field to grow exponentially with depth and cover thousands of timesteps.

At training time, the input sequences are real waveforms recorded from human speakers. After training, we can sample the network to generate synthetic utterances. At each step during sampling a value is drawn from the probability distribution computed by the network. This value is then fed back into the input and a new prediction for the next step is made. Building up samples one step at a time like this is computationally expensive, but we have found it essential for generating complex, realistic-sounding audio.

Improving the State of the Art

We trained WaveNet using some of Google's TTS datasets so we could evaluate its performance. The following figure shows the quality of WaveNets on a scale from 1 to 5, compared with Google's current best TTS systems (**parametric** and **concatenative**), and with human speech using **Mean Opinion Scores (MOS)**. MOS are a standard measure for subjective sound

quality tests, and were obtained in blind tests with human subjects (from over 500 ratings on 100 test sentences). As we can see, WaveNets reduce the gap between the state of the art and human-level performance by over 50% for both US English and Mandarin Chinese.

For both Chinese and English, Google's current TTS systems are considered among the best worldwide, so improving on both with a single model is a major achievement.

Here are some samples from all three systems so you can listen and compare yourself:

US English:

Parametric

Concatenative

WaveNet

Mandarin Chinese:

Parametric

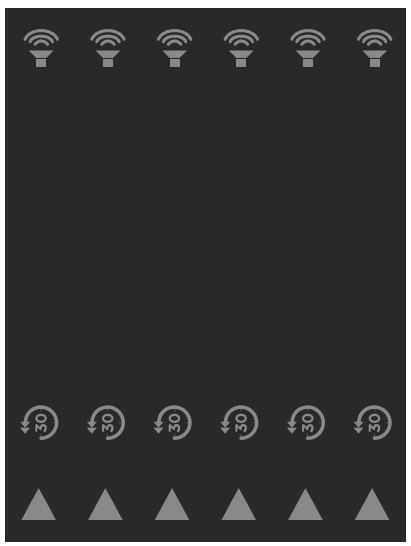
Concatenative

WaveNet

Knowing What to Say

In order to use WaveNet to turn text into speech, we have to tell it what the text is. We do this by transforming the text into a sequence of linguistic and phonetic features (which contain information about the current phoneme, syllable, word, etc.) and by feeding it into WaveNet. This means the network's predictions are conditioned not only on the previous audio samples, but also on the text we want it to say.

If we train the network without the text sequence, it still generates speech, but now it has to make up what to say. As you can hear from the samples below, this results in a kind of babbling, where real words are interspersed with made-up word-like sounds:



DeepMind

Follow

Research

Applied

Blog

About Us

Careers

Press

Terms and Conditions

Privacy Policy

Alphabet Inc