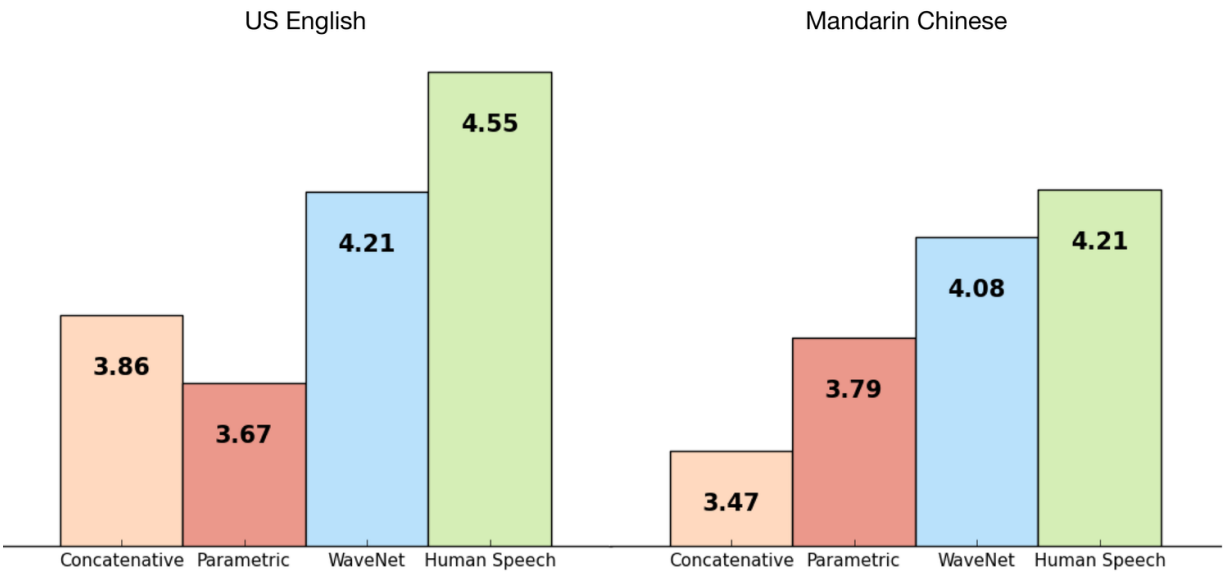
# You may well ask. Who, not what, is talking?

September 12 2016, by Nancy Owano




(Tech Xplore)—This is for real. Human speech synthesis has reached a new high. Thanks to DeepMind, there is every indication that machines are getting quite good at sounding like humans.

Google's DeepMind unit has worked on a system which is earning praise among tech watching sites this month.

Jeremy Kahn at Bloomberg described the DeepMind system as an artificial intelligence called WaveNet that can mimic human speech by

learning how to form the individual sound waves a human voice creates.

The DeepMind team themselves talked about it in a recent blog. WaveNet, they said, is "a deep generative model of raw audio waveforms." WaveNet is said to be directly modelling the raw waveform of the audio signal, one sample at a time.

Ryan Whitwam in *Geek.com* said that it has been difficult to develop text-to-speech (TTS) which sounds authentically human.

*Discover* also talked about how making the reply sound realistic has proven challenging. "Right now, computers are pretty good listeners, because deep learning algorithms have taken speech recognition to a new level. But computers still aren't very good speakers."

"Most TTS systems are based on so-called concatenative technologies. This relies upon a database of speech fragments that are combined to form words. (Carl Engelking in *Discover* referred to it as "basically, cobbling words together from a massive database of sound fragments.") This tends to sound rather uneven and has odd inflections. There is also some work being done on parametric TTS, which uses a data model to generate words, but this sounds even less natural," said Whitwam in *Geek.com*.

What unites the two approaches, said Jamie Condliffe in *MIT Technology Review*, is that "they both stitch together chunks of sound, rather than creating the whole audio waveform from scratch."

Whitwam said the DeepMind approach marks a change in the way speech synthesis is handled—it involves directly modeling the raw waveform of human speech.

The DeepMind post said this:

"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."

Audio generated by WaveNet is more realistic. Condliffe said the results were "noticeably more humanlike" compared with the other two approaches.

How close do the system's soundwaves come to resembling human speech? How humanlike is it? Does it still sound like a robot, but a very humanlike robot?

Kahn said, "In blind tests for U.S. English and Mandarin Chinese, human listeners found WaveNet-generated speech sounded more natural than that created with any of Google's existing text-to-speech programs, which are based on different technologies. WaveNet still underperformed recordings of actual human speech."

They achieved what they did achieve via a neural network. Engelking said, "WaveNet is an artificial neural network, that, at least on paper, resembles the architecture of the human brain."

Engelking in *Discover* looked at the bigger picture. "We're not there yet, but natural language processing is a scorching hot area of AI

research—Amazon, Apple, Google and Microsoft are all in pursuit of savvy digital assistants that can verbally help us interact with our devices." He said "the future of man-machine conversation sounds pretty good."

Similarly, Kahn in Bloomberg made the similar observation: "Speech is becoming an increasingly important way humans interact with everything from mobile phones to cars. Amazon.com Inc., Apple Inc., Microsoft Inc. and Alphabet Inc.'s Google have all invested in personal digital assistants that primarily interact with users through speech."

What's next? "WaveNets open up a lot of possibilities for TTS, music generation and audio modelling in general...We are excited to see what we can do with them next," according to the DeepMind blog.

One thing is clear. As Kahn said, "WaveNet is yet another coup for DeepMind."

G. Clay Whittaker in *Popular Science* meanwhile shared a thought that really is worth thinking about. "Imagine if Siri, Cortana, or Alexa started having inflection, variances, and realistic breathing patterns...So sooner than later, when you hear a voice on a phone, it may be harder to tell if you're hanging up on a telemarketing person or computer." However, he also left off with a compelling thought: "But let's just hope Google's AI doesn't start hearing voices telling it to do things."

**More information:** deepmind.com/blog/wavenet-gene … ive-model-raw-audio/

drive.google.com/file/d/0B3cxc … eWpLVXhkTDJINDQ/view