DAILY NEWS 26 February 2018

Baidu can clone your voice after hearing just a minute of audio

Baidu is building on its Deep Voice engine
Anthony Kwan/Bloomberg via Getty Images

By **Edd Gent**

Chinese search giant Baidu says it can create a copy of someone's voice using neural networks – and all that's needed to work from is less than a minute's worth of audio of the person talking.

Baidu researchers say the technology could create digital duplicate voices for people who have lost the ability to talk. It could also be used to personalise digital assistants, video game characters or automatic speech translation services.

"A mum could easily configure an audio-book reader with her own voice to read bedtime stories for her kids," says Sercan Arik at Baidu Research, who led the work.

Voice cloning technology has improved rapidly in recent years. In 2016 Adobe released VoCo, which could mimic someone's voice using 20 minutes of audio. Last year, Canadian start-up Lyrebird released a service letting anyone create a digital copy of their voice based on 1 minute of audio.

Digital mimic

Baidu's research builds on its text-to speech synthesis system Deep Voice, which was trained on more than 800 hours of audio from 2400 speakers. It builds a model of human speech by learning what sounds go with what text and also learns the idiosyncrasies of each speaker it was trained on.

Now new software is able to synthesise a copy of a voice based on just hearing snatches of the original. The best version needed 100 snippets, each no more than 5 seconds long, the Baidu team says. But one trained on just 10 snippets performed well enough to dupe a voice recognition system more than 95 per cent of the time, and human evaluators gave it 3.16 out of 4 for mimickry.

The output is still not totally indistinguishable from the human voice, says Arik, "but it does show a very fundamental breakthrough in that direction". You can listen to the voices here.

Read more: An AI has learned how to pick a single voice out of a crowd

Rita Singh, a voice forensic science expert at Carnegie Mellon University, points out that even the best synthesised voices contain telltale digital signals easily detected by advanced voice profiling algorithms.

But most voice authentication systems – used to secure everything from banking services to smartphones – can be fooled because they rely instead on picking up broad statistical features, she says.

In 2014, University of Alabama security researcher Nitesh Saxena showed that a freely available voice morphing tool could trick voice authentication systems 80 to 90 per cent of the time. Unpublished research shows leading digital assistants and even a major bank's telephone service remain vulnerable, he says.

But while biometric systems can be improved, our own ability to detect fakes can't. This raises the spectre of voice synthesis systems aping someone's voice to commit fraud or sparking fake news by doctoring a politician's speech.

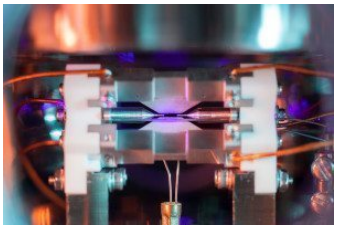
"Humans will over time become even more vulnerable to such attacks," says Saxena.

Reference: arxiv.org/abs/1802.06006

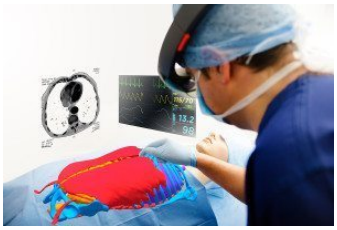
POPULAR

A guide to why your world is a hallucination

The tamed ape: were humans the first animal to be domesticated?



A single atom is visible to the naked eye in this stunning photo



Augmented reality goggles give surgeons X-ray vision

Green is the new black: Redesigning clothes to save the planet

How sharing other people's feelings can make you sick

Bats spread Ebola because they've evolved not to fight viruses

Persistent coughs melt away with chocolate