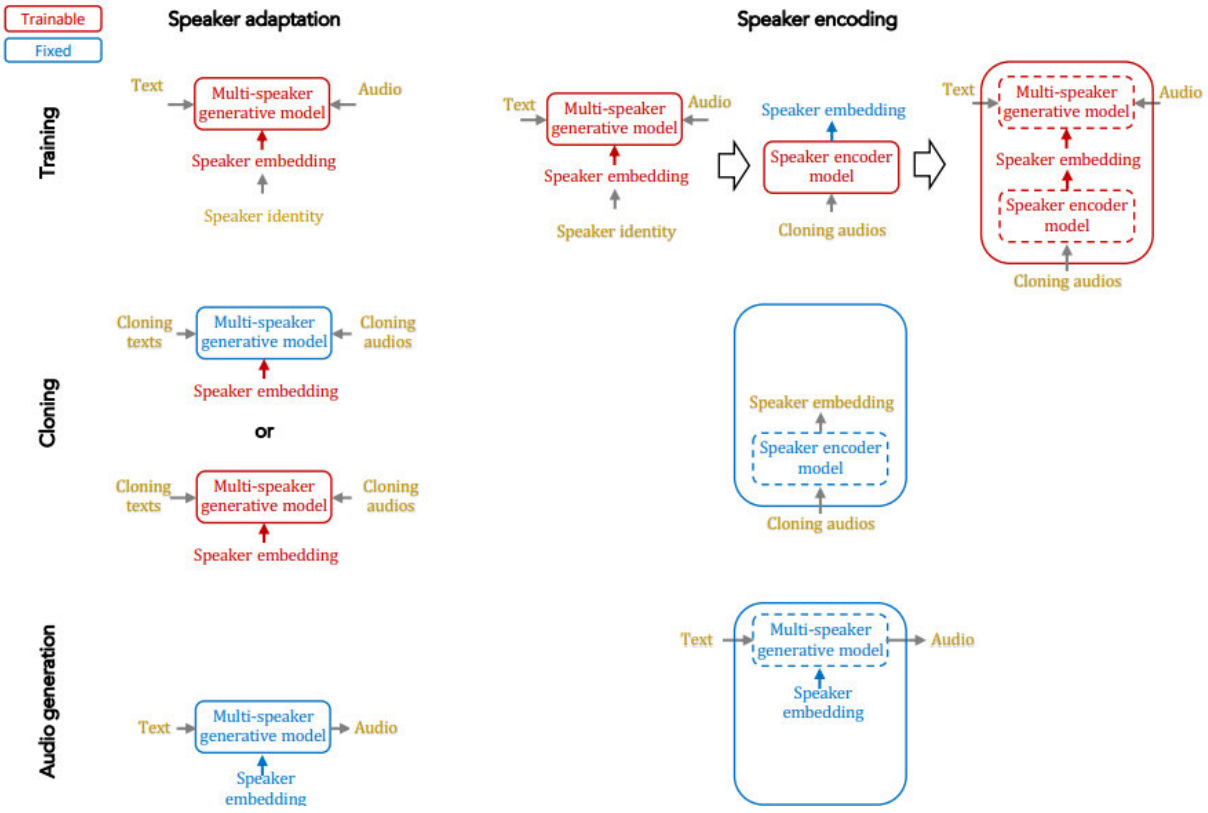
Upgraded Deep Voice can mimic any voice in mere seconds

March 6 2018, by Bob Yirka



Speaker adaptation and speaker encoding approaches for training, cloning and audio generation. Credit: arXiv:1802.06006 [cs.CL]

Via whitepaper which they have uploaded to the *arXiv* preprint server, a team at Baidu (China's answer to Google) has announced an upgrade to their text-to-speech application called Deep Voice. Now, instead of

taking a half-hour or longer to analyze a person's voice and replicate it, the system can do it in less than a minute. The neural-network based system is part of an effort by the team at Baidu to make machines sound more like humans when they "speak" to us.

There are two parts to the system. The first involves recording voice samples to allow the system to learn what the subject's voice sounds like. The second part reads user-defined text aloud in the voice of the subject.

Several groups have been working on projects aimed at replicating the sound of an individual person's voice, ostensibly to allow robot assistants to sound like actual human assistants. Thus, a program that converts text into words that sounds like you, your neighbor, Donald Trump or the Queen of England is not expected to offer much in the way of an end product—though Baidu does suggest it could be used by people who have lost the use of their voice. Instead, it is meant as a stepping stone to greater things. The new system, the team reports, works optimally when given 100 five-second voice samples. It can also manipulate a voice, allowing people to hear how they might sound, for example, with a British accent, or as someone of the opposite gender. It is also getting better at mimicking voices, and is now able to fool voice recognition software 95 percent of the time—and a human test gave the system an average rating of 3.16 out of 4.

But, as many in the press have noted, the technology could cause problems. Taped interrogations by police could become useless if anyone with a smartphone could generate the same conversation. There is also the problem of identity theft. If a thief can steal your data and your voice , you might never get it back. Or consider political operatives releasing fake recordings of politicians having conversations that could sway an election.

More information: — Baidu Research blog:

research.baidu.com/neural-voice-cloning-samples/

— Samples: audiodemos.github.io/

— Neural Voice Cloning with a Few Samples, arXiv:1802.06006 [cs.CL] arxiv.org/abs/1802.06006

Abstract
Voice cloning is a highly desired feature for personalized speech interfaces. Neural network based speech synthesis has been shown to generate high quality speech for a large number of speakers. In this paper, we introduce a neural voice cloning system that takes a few audio samples as input. We study two approaches: speaker adaptation and speaker encoding. Speaker adaptation is based on fine-tuning a multi-speaker generative model with a few cloning samples. Speaker encoding is based on training a separate model to directly infer a new speaker embedding from cloning audios and to be used with a multi-speaker generative model. In terms of naturalness of the speech and its similarity to original speaker, both approaches can achieve good performance, even with very few cloning audios. While speaker adaptation can achieve better naturalness and similarity, the cloning time or required memory for the speaker encoding approach is significantly less, making it favorable for low-resource deployment.

Citation: Upgraded Deep Voice can mimic any voice in mere seconds (2018, March 6) retrieved 1 May 2024 from https://techxplore.com/news/2018-03-deep-voice-mimic-mere-seconds.html