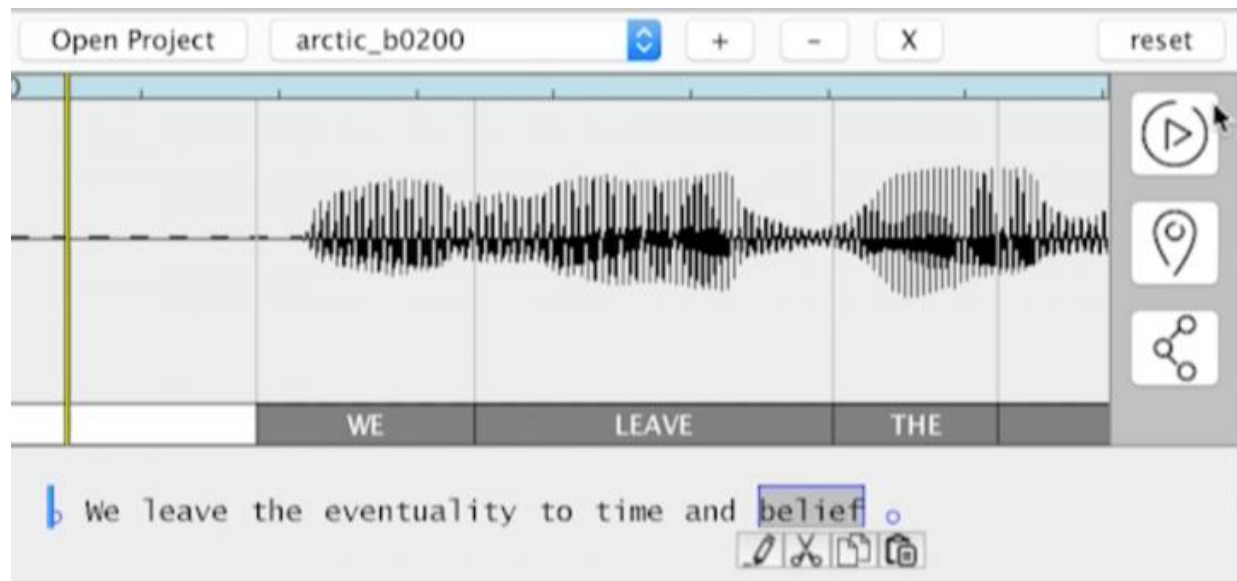# Technology edits voices like text

May 15 2017

Technology invented by Princeton University computer scientists allows people to edit audio recordings with the ease of changing words on a computer screen. The system inserts new words in the same voice as the rest of the recording. Credit: Professor Adam Finkelstein

Anyone who ever used a typewriter will recall the difficulty of fixing a misspelled or poorly chosen word—remember whiteout and correction tape?

Now, technology developed by Princeton University computer scientists may do for audio recordings of the human voice what word processing software did for the written word.

The software, named VoCo, provides an easy means to add or replace a word in an audio recording of a human voice by editing a transcript of the recording. New words are automatically synthesized in the speaker's voice even if they don't appear anywhere else in the recording.

The system, which uses a sophisticated algorithm to learn and recreate the sound of a particular voice, could one day make editing podcasts and narration in videos much easier. More broadly, the technology could provide a launching point for creating personalized robotic voices that sound natural.

"VoCo provides a peek at a very practical technology for editing audio tracks, but it is also a harbinger for future technologies that will allow the human voice to be synthesized and automated in remarkable ways," said Adam Finkelstein, a professor of computer science at Princeton.

Zeyu Jin, a Princeton graduate student advised by Finkelstein, will present the work at the Association for Computing Machinery SIGGRAPH conference in July. The work at Princeton was funded by the Project X Fund, which provides seed funding to engineers for pursuing speculative projects. The Princeton researchers collaborated with scientists Gautham Mysore, Stephen DiVerdi, and Jingwan Lu at Adobe Research.

The team described the development of VoCo in a paper to be published in the July issue of the journal *Transactions on Graphics*.

On a computer screen, VoCo's user interface looks similar to other audio editing software such as the popular podcast editing program Audacity or Apple's music editing program GarageBand. It offers visualization of the waveform of the audio track and a set of cut, copy and paste tools for editing. Unlike other programs, however, VoCo also augments the waveform with a text transcript of the track and allows the user to

replace or insert new words that don't already exist in the track simply by typing in the transcript. When the user types the new word, VoCo updates the audio track, automatically synthesizing the new word by stitching together snippets of audio from elsewhere in the narration.

"Currently, audio editors can cut out pieces of a track of narration and move a clip from one place to another. However, if you want to add a word that doesn't exist in the recording, it's possible only through a painstaking trial and error process of searching for small audio snippets that might fit together well enough to plausibly form the word," said Finkelstein. "VoCo automates the search and stitching process, and produces results that typically sound even better than those created manually by audio experts."

At the heart of VoCo is an optimization algorithm that searches the voice recording and chooses the best possible combinations of partial word sounds, called "phonemes," to build new words in the user's voice. To do this, it not only needs to find the individual phonemes, but also find sequences of them that stitch together without abrupt transitions, as well as fit them into the existing sentence so that the new word blends in seamlessly. Words are pronounced with different emphasis and intonation depending on where they fall in a sentence, so context is important.

For clues about this context, VoCo looks to an audio track of the sentence that is automatically synthesized in artificial voice from the text transcript—one that sounds robotic to human ears. This recording is used as a point of reference in building the new word. VoCo then matches the pieces of sound from the real human voice recording to match the word in the synthesized track—a technique known as "voice conversion," which inspired the project name VoCo.

In case the synthesized word isn't quite right, VoCo offers users several

versions of the word to choose from. The system also provides an advanced editor to modify pitch and duration, allowing expert users to further polish the track.

To test how effective their system was a producing authentic sounding edits, the researchers asked people to listen to a set of audio tracks, some of which had been edited with VoCo and other that were completely natural. The fully automated versions were mistaken for real recordings more than 60 percent of the time.

Jin, whose research interests straddle audio and machine learning, said voice conversion technologies hold promise for a range of applications beyond editing audio tracks. For instance, people who have lost their voices due to injury or disease might be able to recreate their voices through a robotic system.

"We were approached by a man who has a neurodegenerative disease and can only speak through a text to speech system controlled by his eyelids," said Jin. "The voice sounds robotic, like the system used by Steven Hawking, but he wants his young daughter to hear his real voice. It might one day be possible to analyze past recordings of him speaking and created an assistive device that speaks in his own voice."

On the lighter side, Jin said voice conversion might be used to bring back the long lost voices of iconic cartoon characters such as Bugs Bunny or Popeye. Such voices—and those of famous actors or historic figures—could then be used to create narration for new movies, or even integrated into automated intelligent personal assistants like Apple's Siri or Amazon's Alexa.

The Princeton researchers are currently refining the VoCo algorithm to improve the system's ability to integrated synthesized words more smoothly into audio tracks. They are also working to expand the system's

capabilities to create longer phrases or even entire sentences synthesized from a narrator's [voice](#).

Finkelstein said that editing software like VoCo raises important questions about how to treat digital content when we know it may have been altered to change its meaning. "This question came to the forefront for photography decades ago with the arrival of digital image editing software like Adobe Photoshop," he said.

He said the emergence of fast and easy photo editing led to long discussions of the reliability of photos in news stories. Even before digital editing became available, expert photographers had many tricks for modifying their prints, but new programs made it faster and easier, and did not require the same degree of expertise.

"Today we take it for granted that photos can be edited, and we judge photos with a little more skepticism," he said. "We understand there is a journalistic responsibility attached to photos."

He said the same discussion is now happening with digital audio. Editors have long been able to modify audio files to clean up an audio track, and they could choose to change its meaning, for example simply by removing the word "not." But he said that programs like VoCo, by making that process easier, will likely raise concerns.

"This tool will almost certainly fuel the conversation about audio that was preceded by a conversation about photos," Finkelstein said. "Soon enough, it will be followed by a conversation about video."

  More information: The research team has posted preprint of the paper as well as a video demonstrating the project and examples of synthesized voices at gfx.cs.princeton.edu/pubs/Jin_2017_VTI/