Lip-syncing Obama: New tools turn audio clips into realistic video

July 11 2017, by Jennifer Langston

Credit: University of Washington

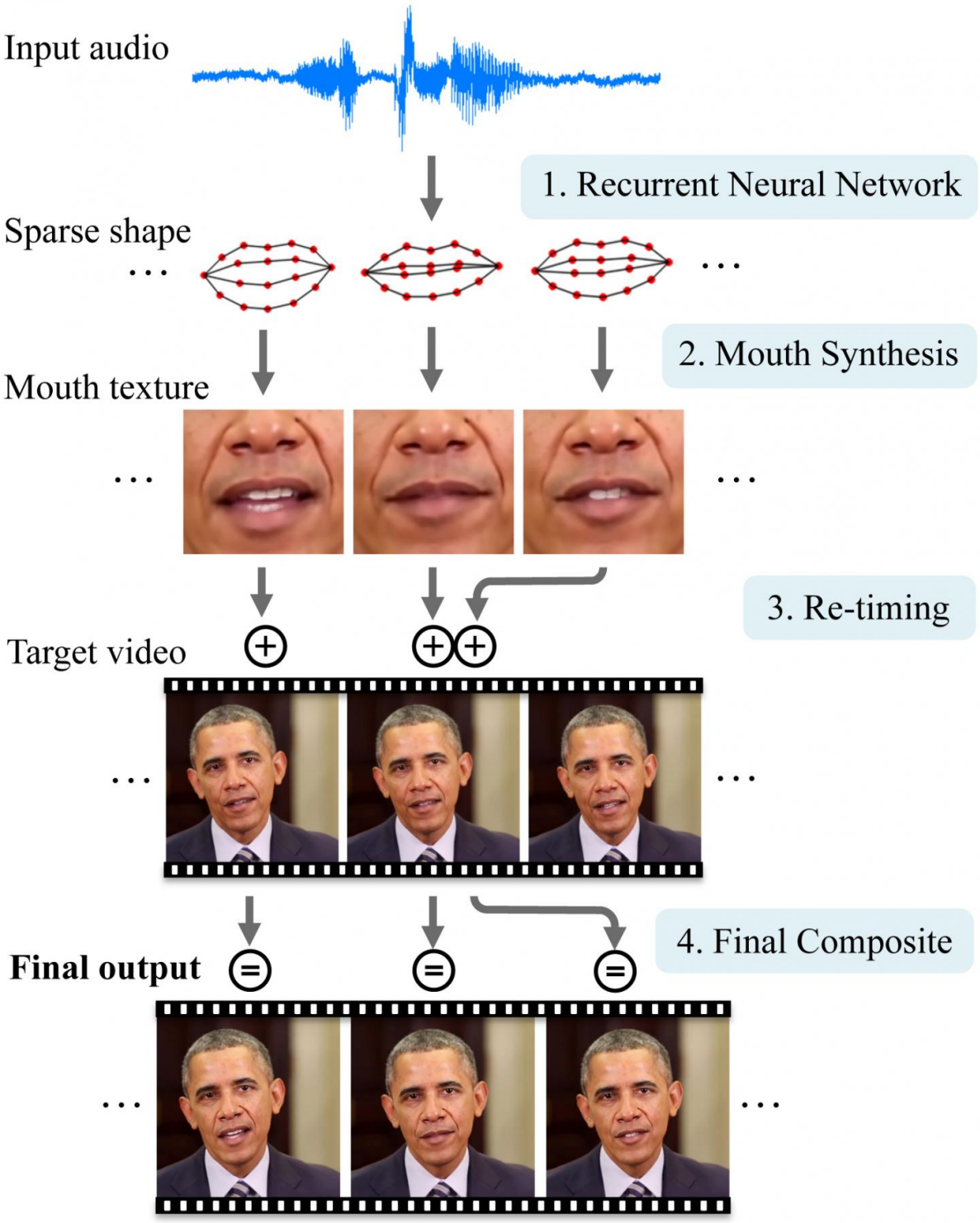
University of Washington researchers have developed new algorithms that solve a thorny challenge in the field of computer vision: turning audio clips into a realistic, lip-synced video of the person speaking those words.

As detailed in a paper to be presented Aug. 2 at SIGGRAPH 2017, the team successfully generated highly-realistic video of former president Barack Obama talking about terrorism, fatherhood, job creation and other topics using audio clips of those speeches and existing weekly video addresses that were originally on a different topic.

"These type of results have never been shown before," said Ira

Kemelmacher- Shlizerman, an assistant professor at the UW's Paul G. Allen School of Computer Science & Engineering. "Realistic audio-to-video conversion has practical applications like improving video conferencing for meetings, as well as futuristic ones such as being able to hold a conversation with a historical figure in virtual reality by creating visuals just from audio. This is the kind of breakthrough that will help enable those next steps."

In a visual form of lip-syncing, the system converts audio files of an individual's speech into realistic mouth shapes, which are then grafted onto and blended with the head of that person from another existing video.



A neural network first converts the sounds from an audio file into basic mouth shapes. Then the system grafts and blends those mouth shapes onto an existing

target video and adjusts the timing to create a realistic, lip-synced video of the person delivering the new speech. Credit: University of Washington

The team chose Obama because the machine learning technique needs available video of the person to learn from, and there were hours of presidential videos in the public domain. "In the future video, chat tools like Skype or Messenger will enable anyone to collect videos that could be used to train computer models," Kemelmacher-Shlizerman said.

Because streaming audio over the internet takes up far less bandwidth than video, the new system has the potential to end video chats that are constantly timing out from poor connections.

"When you watch Skype or Google Hangouts, often the connection is stuttery and low-resolution and really unpleasant, but often the audio is pretty good," said co-author and Allen School professor Steve Seitz. "So if you could use the audio to produce much higher-quality video, that would be terrific."

By reversing the process—feeding video into the network instead of just audio—the team could also potentially develop algorithms that could detect whether a video is real or manufactured.

The new machine learning tool makes significant progress in overcoming what's known as the "uncanny valley" problem, which has dogged efforts to create realistic video from audio. When synthesized human likenesses appear to be almost real—but still manage to somehow miss the mark—people find them creepy or off-putting.

"People are particularly sensitive to any areas of your mouth that don't look realistic," said lead author Supasorn Suwajanakorn, a recent

doctoral graduate in the Allen School. "If you don't render teeth right or the chin moves at the wrong time, people can spot it right away and it's going to look fake. So you have to render the mouth region perfectly to get beyond the uncanny valley."

Previously, audio-to-video conversion processes have involved filming multiple people in a studio saying the same sentences over and over to try to capture how a particular sound correlates to different mouth shapes, which is expensive, tedious and time-consuming. By contrast, Suwajanakorn developed algorithms that can learn from videos that exist "in the wild" on the internet or elsewhere.

"There are millions of hours of video that already exist from interviews, video chats, movies, television programs and other sources. And these deep learning algorithms are very data hungry, so it's a good match to do it this way," Suwajanakorn said.

Rather than synthesizing the final video directly from audio, the team tackled the problem in two steps. The first involved training a neural network to watch videos of an individual and translate different audio sounds into basic mouth shapes.

By combining previous research from the UW Graphics and Image Laboratory team with a new mouth synthesis technique, they were then able to realistically superimpose and blend those mouth shapes and textures on an existing reference video of that person. Another key insight was to allow a small time shift to enable the neural network to anticipate what the speaker is going to say next.

The new lip-syncing process enabled the researchers to create realistic videos of Obama speaking in the White House, using words he spoke on a television talk show or during an interview decades ago.

Currently, the neural network is designed to learn on one individual at a time, meaning that Obama's voice—speaking words he actually uttered—is the only information used to "drive" the synthesized video. Future steps, however, include helping the algorithms generalize across situations to recognize a person's voice and speech patterns with less data - with only an hour of video to learn from, for instance, instead of 14 hours.

"You can't just take anyone's voice and turn it into an Obama video," Seitz said. "We very consciously decided against going down the path of putting other people's words into someone's mouth. We're simply taking real words that someone spoke and turning them into realistic video of that individual."

Provided by University of Washington