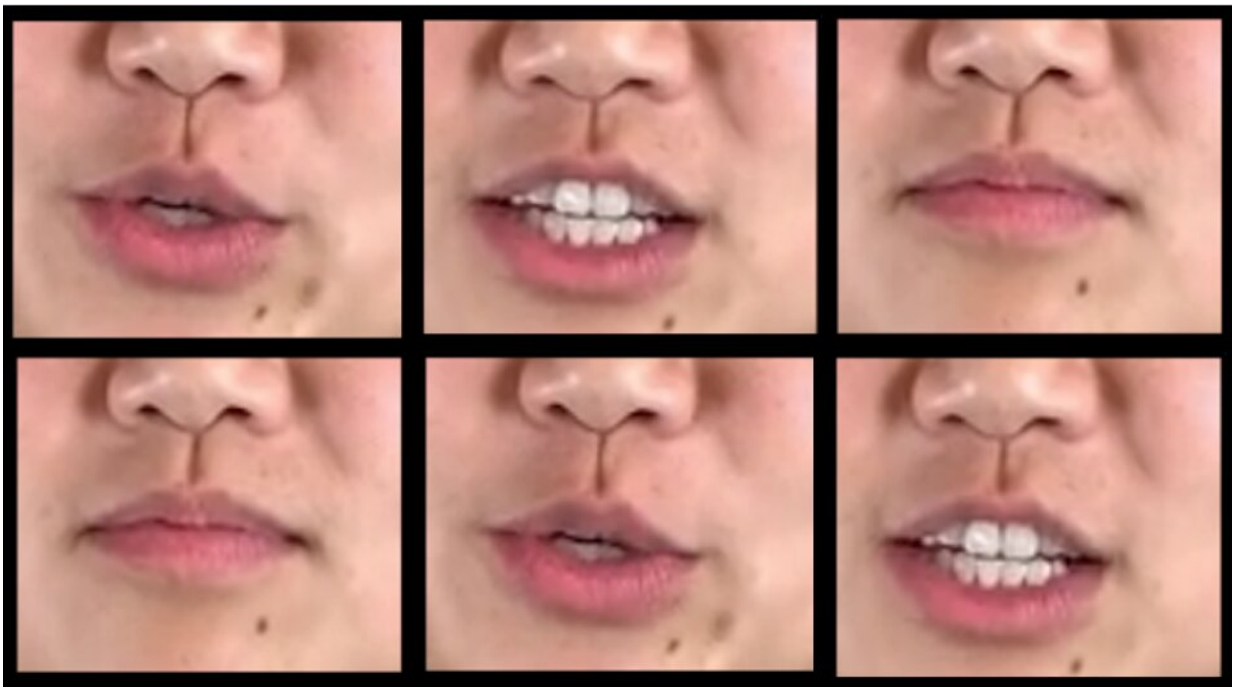
Using AI to detect seemingly perfect deep-fake videos

October 14 2020, by Edmund L. Andrews



To spot a deep fake, researchers looked for inconsistencies between “visemes,” or mouth formations, and “phonemes,” the phonetic sounds. Credit: Stanford University

One year ago, Maneesh Agrawala of Stanford helped develop a lip-sync technology that allowed video editors to almost undetectably modify speakers' words. The tool could seamlessly insert words that a person never said, even mid-sentence, or eliminate words she had said. To the

naked eye, and even to many computer-based systems, nothing would look amiss.

The tool made it much easier to fix glitches without re-shooting entire scenes, as well as to tailor TV shows or movies for different audiences in different places.

But the technology also created worrisome new opportunities for hard-to-spot deep-fake videos that are created for the express purpose of distorting the truth. A recent Republican video, for example, used a cruder technique to doctor an interview with Vice President Joe Biden.

This summer, Agrawala and colleagues at Stanford and UC Berkeley unveiled an AI-based approach to detect the lip-sync technology. The new program accurately spots more than 80 percent of fakes by recognizing minute mismatches between the sounds people make and the shapes of their mouths.

But Agrawala, the director of Stanford's Brown Institute for Media Innovation and the Forest Baskett Professor of Computer Science, who is also affiliated with the Stanford Institute of Human-Centered Artificial Intelligence, warns that there is no long-term technical solution to deep fakes.

The real task, he says, is to increase media literacy to hold people more accountable if they deliberately produce and spread misinformation.

"As the technology to manipulate video gets better and better, the capability of technology to detect manipulation will get worse and worse," he says. "We need to focus on non-technical ways to identify and reduce disinformation and misinformation."

The manipulated video of Biden, for example, was exposed not by the

technology but rather because the person who had interviewed the vice president recognized that his own question had been changed.

How deep fakes work

There are legitimate reasons for manipulating video. Anyone producing a fictional TV show, a movie or a commercial, for example, can save time and money by using digital tools to clean up mistakes or tweak scripts.

The problem comes when those tools are intentionally used to spread false information. And many of the techniques are invisible to ordinary viewers.

Many deep-fake videos rely on face-swapping, literally super-imposing one person's face over the video of someone else. But while face-swapping tools can be convincing, they are relatively crude and usually leave digital or visual artifacts that a computer can detect.

Lip-sync technologies, on the other hand, are more subtle and thus harder to spot. They manipulate a much smaller part of the image, and then synthesize lip movements that closely match the way a person's mouth really would have moved if he or she had said particular words. With enough samples of a person's image and voice, says Agrawala, a deep-fake producer can get a person to "say" anything.

Spotting the fakes

Worried about unethical uses of such technology, Agrawala teamed up on a detection tool with Ohad Fried, a postdoctoral fellow at Stanford; Hany Farid, a professor at UC Berkeley's School of Information; and Shruti Agarwal, a doctoral student at Berkeley.

The basic idea is to look for inconsistencies between "visemes," or mouth formations, and "phonemes," the phonetic sounds. Specifically, the researchers looked at the person's mouth when making the sounds of a "B," "M," or "P," because it's almost impossible to make those sounds without firmly closing the lips.

The researchers first experimented with a purely manual technique, in which human observers studied frames of video. That worked well but was both labor-intensive and time-consuming in practice.

The researchers then tested an AI-based neural network, which would be much faster, to make the same analysis after training it on videos of former President Barack Obama. The neural network spotted well over 90 percent of lip-syncs involving Obama himself, though the accuracy dropped to about 81 percent in spotting them for other speakers.

A real truth test

The researchers say their approach is merely part of a "cat-and-mouse" game. As deep-fake techniques improve, they will leave even fewer clues behind.

In the long run, Agrawala says, the real challenge is less about fighting deep-fake videos than about fighting disinformation. Indeed, he notes, most disinformation comes from distorting the meaning of things people actually have said.

"Detecting whether a video has been manipulated is different from detecting whether the video contains misinformation or disinformation, and the latter is much, much harder," says Agrawala.

"To reduce disinformation, we need to increase media literacy and develop systems of accountability," he says. "That could mean laws

against deliberately producing disinformation and consequences for breaking them, as well as mechanisms to repair the harms caused as a result."

More information: Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. www.ohadf.com/papers/AgarwalFa … rawala_CVPRW2020.pdf

Provided by Stanford University