“Video Research at MIT Puts Words into Mouths, with Startling Results”

By THEO EMERY, Associated Press Writer
Sun Jun 30, 2:47 AM ET

CAMBRIDGE, Mass. - Marilyn Monroe died a generation before karaoke, digital animation and the pop singer Dido's race onto Billboard's Top Ten.

But in an eerie video clip created by Massachusetts Institute of Technology (news - web sites) researchers, the long-dead celluloid star croons the song "Hunter" by the very much alive performer.

The MIT team has combined artificial intelligence and videography to make words and song — even in foreign languages — emerge from the lips of people who could never possibly have uttered them.

"We wanted to try it on a celebrity who wasn't alive," said Tony F. Ezzat, an MIT postdoctoral fellow who created the surreal sequence. "I'm thrilled, obviously."

Yet not only does the video blur past and present. It also heralds new possibilities for video mischief. Just as digital stills can be manipulated to misrepresent reality, so will advances in digital video technology enable full-motion fakery.

The new video sleight of hand is the work of a group headed by Tomaso Poggio, a professor at MIT's McGovern Institute and Artificial Intelligence Lab whose efforts in the early 1990s involved cartoons. Poggio and others later created 3-D images of faces that could be viewed from a variety of angles with a range of emotional expressions.

Now, Poggio and Ezzat have programmed a computer to troll through short video clips and learn how a specific person speaks, a process that can take several days.

Once the computer has learned how the person shapes their mouth around individual sound segments — called "phonemes" — it can digitally morph the shape of the subject's mouth around any audio sequence the creator wants to put words in a subject's mouth within minutes.

It's the "teaching" of the computer that makes this method different from most existing facial animation technology.

The recorded results? A woman made to sing in Japanese, and Marilyn Monroe lip-synching a song that didn't become famous until decades after her death. Ezzat has also been working on a video of Ted Koppel, ABC's "Nightline" anchor, speaking in Spanish.

The MIT team is most excited to see this new technology used for language training, helping the deaf learn to speak or putting a more human face on computers, though it also has obvious applications for entertainment and film, such as realistic dubbing.

Bob Steele, director of the ethics program at The Poynter Institute, a journalism research center, worries about the potential for abuse.

There are serious concerns that videotape could be doctored for unethical purposes: to fabricate evidence and literally put words in someone's mouth.

Consider, for example, all the tapes — such as those from the Nixon presidency — that have recently been released. How will the future authenticity of such audio recordings be guaranteed?

"If we use this new technology in a way that can alter reality, we certainly run the risk of deceiving the public, and if we deceive the public, we individually and collectively increase their skepticism of what they read in the paper and magazines, see on television or hear on the radio, or read online," Steele said.

Advances in digital imaging have already forced newspapers to adopt guidelines and protocols for altering photographs. Similar measures will need to be taken for facial animation, said Steele.

There is still time to mull policy on the technology's use. At this stage, it remains far from perfect.

In the MIT videos, little facial expression accompanies the animated mouth movements, and extended viewing reveals that rest of the face isn't always in synch with the mouth.

In addition, the technology's success depends on the subject staring straight ahead, without much head movement.

A study by Gadi Geiger, who authored the MIT research paper with Poggio and Ezzat, found that 22 adult viewers were only able to discern the real images from the animated images about half the time.

The MIT team does worry that its work, to be presented next month in San Antonio at Siggraph, might be used for mischief or worse. But its members say there's no more danger of misuse than with any other innovation.

"It's a worry with any kind of research. This, I think, is not the one to worry about," Poggio said.

Ezzat said video "watermarks" could prevent copying or alteration of tapes in the same way treasury bills are marked to prevent forgery.

Also, he said, the computer can't fabricate words; it only superimposes real sound onto video, so — as of now — it would be difficult to make people say things on the MIT videotapes that they never actually said.

Bill DeStefanis, a vice president at the imaging and speech recognition software company Scan Soft Inc., said it would be very difficult to meld computer-generated speech and animated video realistically enough to fool anyone.

"The technology is not good enough to trick a human ear," he said. "We do have the ability to take recordings of someone, and build a synthesized voice and get a machine to say something that sounds like them, but you can tell the difference."