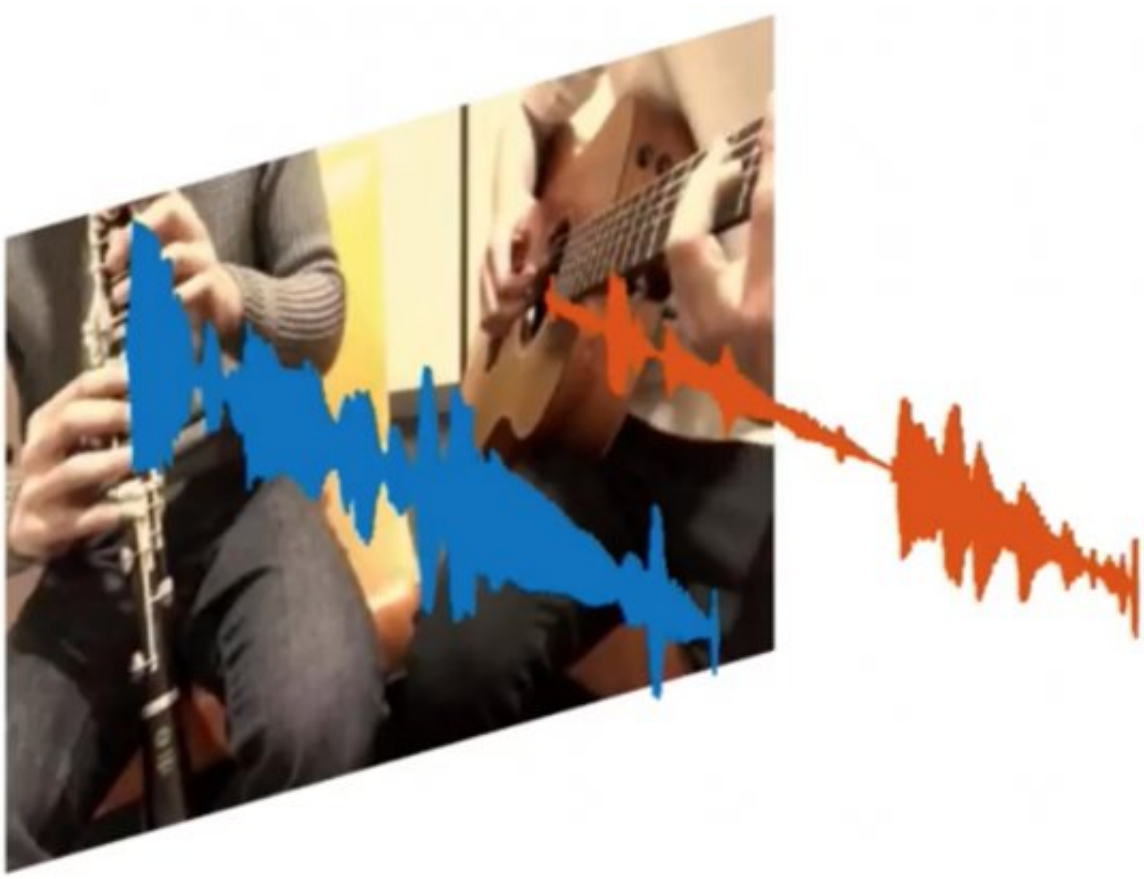
An AI system for editing music in videos

July 6 2018, by Adam Conner-Simons



A new AI system called PixelPlayer can look at an image and determine which set of pixels are responsible for making specific sets of soundwaves. Credit: MIT CSAIL

Amateur and professional musicians alike may spend hours pouring over

YouTube clips to figure out exactly how to play certain parts of their favorite songs. But what if there were a way to play a video and isolate the only instrument you wanted to hear?

That's the outcome of a new AI project out of MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL): a deep-learning system that can look at a video of a musical performance, and isolate the sounds of specific instruments and make them louder or softer.

The system, which is "self-supervised," doesn't require any human annotations on what the instruments are or what they sound like.

Trained on over 60 hours of videos, the "PixelPlayer" system can view a never-before-seen musical performance, identify specific instruments at pixel level, and extract the sounds that are associated with those instruments.

For example, it can take a video of a tuba and a trumpet playing the "Super Mario Brothers" theme song, and separate out the soundwaves associated with each instrument.

The researchers say that the ability to change the volume of individual instruments means that in the future, systems like this could potentially help engineers improve the audio quality of old concert footage. You could even imagine producers taking specific instrument parts and previewing what they would sound like with other instruments (i.e. an electric guitar swapped in for an acoustic one).

In a new paper, the team demonstrated that PixelPlayer can identify the sounds of more than 20 commonly seen instruments. Lead author Hang Zhao says that the system would be able to identify many more instruments if it had more training data, though it still may have trouble handling subtle differences between subclasses of instruments (such as

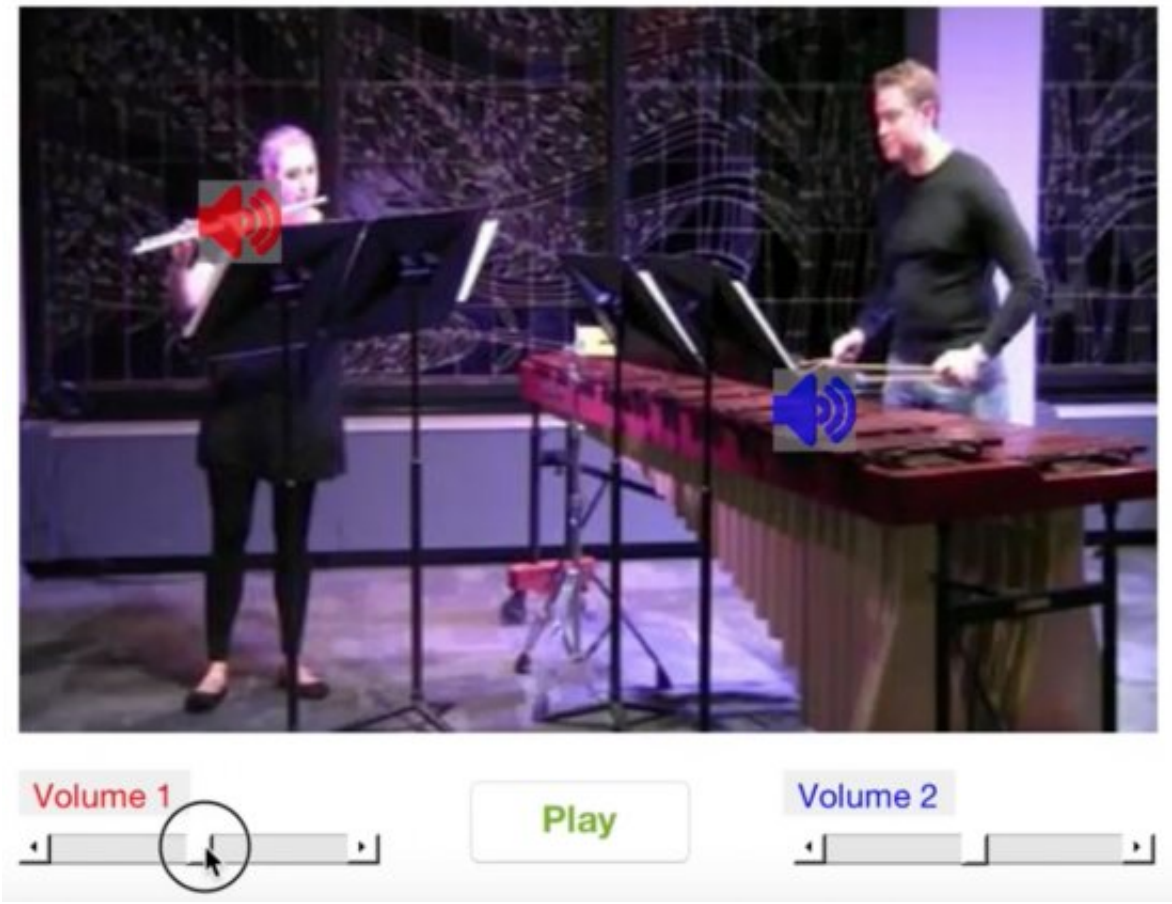
an alto sax versus a tenor).

Previous efforts to separate the sources of sound have focused exclusively on audio, which often requires extensive human labeling. In contrast, PixelPlayer introduces the element of vision, which the researchers say makes human labels unnecessary, as vision provides self-supervision.

The system first locates the image regions that produce sounds, and then separates the input sounds into a set of components that represent the sound from each pixel.

"We expected a best-case scenario where we could recognize which instruments make which kinds of sounds," says Zhao, a Ph.D. student at CSAIL. "We were surprised that we could actually spatially locate the instruments at the pixel level. Being able to do that opens up a lot of possibilities, like being able to edit the audio of individual instruments by a single click on the video."

PixelPlayer uses methods of "deep learning," meaning that it finds patterns in data using so-called "neural networks" that have been trained on existing videos. Specifically, one neural network analyzes the visuals of the video, one analyzes the audio, and a third "synthesizer" associates specific pixels with specific soundwaves to separate the different sounds.



PixelPlayer also includes an interface that lets users change the volume of specific instruments in the mix. Credit: MIT CSAIL

The fact that PixelPlayer uses so-called "self-supervised" deep learning means that the MIT team doesn't explicitly understand every aspect of how it learns which instruments make which sounds.

However, Zhao says that he can tell that the system seems to recognize actual elements of the music. For example, certain harmonic frequencies seem to correlate to instruments like violin, while quick pulse-like patterns correspond to instruments like the xylophone.

Zhao says that a system like PixelPlayer could even be used on robots to

better understand the environmental sounds that other objects make, such as animals or vehicles.

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology