Algorithm recovers speech from vibrations of potato-chip bag filmed through soundproof glass

August 4 2014, by Larry Hardesty

Credit: Christine Daniloff/MIT

Researchers at MIT, Microsoft, and Adobe have developed an algorithm that can reconstruct an audio signal by analyzing minute vibrations of objects depicted in video. In one set of experiments, they were able to recover intelligible speech from the vibrations of a potato-chip bag photographed from 15 feet away through soundproof glass.

In other experiments, they extracted useful audio signals from videos of aluminum foil, the surface of a glass of water, and even the leaves of a potted plant. The researchers will present their findings in a paper at this year's Siggraph, the premier computer graphics conference.

"When sound hits an object, it causes the object to vibrate," says Abe Davis, a graduate student in electrical engineering and computer science at MIT and first author on the new paper. "The motion of this vibration creates a very subtle visual signal that's usually invisible to the naked eye. People didn't realize that this information was there."

Joining Davis on the Siggraph paper are Frédo Durand and Bill Freeman, both MIT professors of computer science and engineering; Neal Wadhwa, a graduate student in Freeman's group; Michael Rubinstein of Microsoft Research, who did his PhD with Freeman; and Gautham Mysore of Adobe Research.

Reconstructing audio from video requires that the frequency of the video samples—the number of frames of video captured per second—be higher than the frequency of the audio signal. In some of their experiments, the researchers used a high-speed camera that captured 2,000 to 6,000 frames per second. That's much faster than the 60 frames per second possible with some smartphones, but well below the frame rates of the best commercial high-speed cameras, which can top 100,000 frames per second.

Commodity hardware

In other experiments, however, they used an ordinary digital camera. Because of a quirk in the design of most cameras' sensors, the researchers were able to infer information about high-frequency vibrations even from video recorded at a standard 60 frames per second. While this audio reconstruction wasn't as faithful as it was with the high-

speed camera, it may still be good enough to identify the gender of a speaker in a room; the number of speakers; and even, given accurate enough information about the acoustic properties of speakers' voices, their identities.

The researchers' technique has obvious applications in law enforcement and forensics, but Davis is more enthusiastic about the possibility of what he describes as a "new kind of imaging."

"We're recovering sounds from objects," he says. "That gives us a lot of information about the sound that's going on around the object, but it also gives us a lot of information about the object itself, because different objects are going to respond to sound in different ways." In ongoing work, the researchers have begun trying to determine material and structural properties of objects from their visible response to short bursts of sound.

In the experiments reported in the Siggraph paper, the researchers also measured the mechanical properties of the objects they were filming and determined that the motions they were measuring were about a tenth of micrometer. That corresponds to five thousandths of a pixel in a close-up image, but from the change of a single pixel's color value over time, it's possible to infer motions smaller than a pixel.

Suppose, for instance, that an image has a clear boundary between two regions: Everything on one side of the boundary is blue; everything on the other is red. But at the boundary itself, the camera's sensor receives both red and blue light, so it averages them out to produce purple. If, over successive frames of video, the blue region encroaches into the red region—even less than the width of a pixel—the purple will grow slightly bluer. That color shift contains information about the degree of encroachment.

Putting it together

Some boundaries in an image are fuzzier than a single pixel in width, however. So the researchers borrowed a technique from earlier work on algorithms that amplify minuscule variations in video, making visible previously undetectable motions: the breathing of an infant in the neonatal ward of a hospital, or the pulse in a subject's wrist.

That technique passes successive frames of video through a battery of image filters, which are used to measure fluctuations, such as the changing color values at boundaries, at several different orientations—say, horizontal, vertical, and diagonal—and several different scales.

The researchers developed an algorithm that combines the output of the filters to infer the motions of an object as a whole when it's struck by sound waves. Different edges of the object may be moving in different directions, so the algorithm first aligns all the measurements so that they won't cancel each other out. And it gives greater weight to measurements made at very distinct edges—clear boundaries between different color values.

The researchers also produced a variation on the algorithm for analyzing conventional video. The sensor of a digital camera consists of an array of photodetectors—millions of them, even in commodity devices. As it turns out, it's less expensive to design the sensor hardware so that it reads off the measurements of one row of photodetectors at a time. Ordinarily, that's not a problem, but with fast-moving objects, it can lead to odd visual artifacts. An object—say, the rotor of a helicopter—may actually move detectably between the reading of one row and the reading of the next.

For Davis and his colleagues, this bug is a feature. Slight distortions of

the edges of objects in conventional video, though invisible to the naked eye, contain information about the objects' high-frequency vibration. And that information is enough to yield a murky but potentially useful audio signal.

"This is new and refreshing. It's the kind of stuff that no other group would do right now," says Alexei Efros, an associate professor of electrical engineering and computer science at the University of California at Berkeley. "We're scientists, and sometimes we watch these movies, like James Bond, and we think, 'This is Hollywood theatrics. It's not possible to do that. This is ridiculous.' And suddenly, there you have it. This is totally out of some Hollywood thriller. You know that the killer has admitted his guilt because there's surveillance footage of his potato chip bag vibrating."

Efros agrees that the characterization of material properties could be a fruitful application of the technology. But, he adds, "I'm sure there will be applications that nobody will expect. I think the hallmark of good science is when you do something just because it's cool and then somebody turns around and uses it for something you never imagined. It's really nice to have this type of creative stuff."

More information: "The visual microphone: passive recovery of sound from video." Abe Davis, et al. *Journal ACM Transactions on Graphics* (*TOG*), Volume 33 Issue 4, July 2014, Article No. 79. DOI: 10.1145/2601097.2601119

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