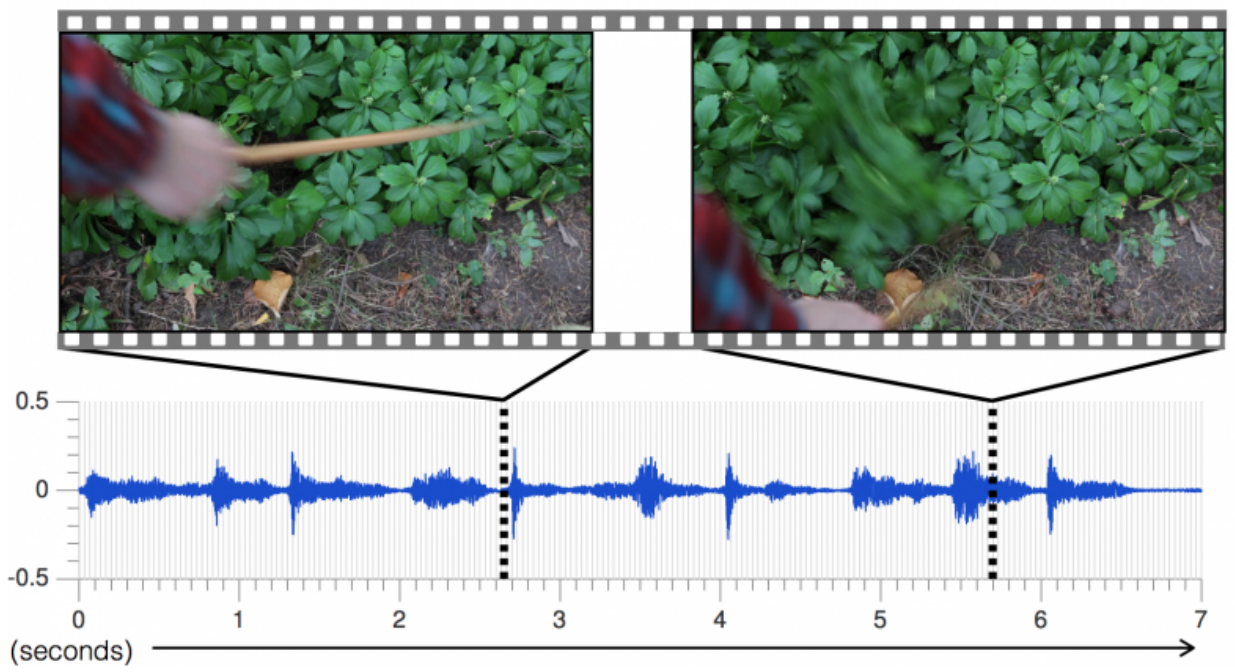
Artificial intelligence produces realistic sounds that fool humans

June 13 2016, by Adam Conner-Simons



Credit: Massachusetts Institute of Technology

For robots to navigate the world, they need to be able to make reasonable assumptions about their surroundings and what might happen during a sequence of events.

One way that humans come to learn these things is through sound. For infants, poking and prodding objects is not just fun; some studies suggest

that it's actually how they develop an intuitive theory of physics. Could it be that we can get machines to learn the same way?

Researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have demonstrated an algorithm that has effectively learned how to predict sound: When shown a silent video clip of an object being hit, the algorithm can produce a sound for the hit that is realistic enough to fool human viewers.

This "Turing Test for sound" represents much more than just a clever computer trick: Researchers envision future versions of similar algorithms being used to automatically produce sound effects for movies and TV shows, as well as to help robots better understand objects' properties.

"When you run your finger across a wine glass, the sound it makes reflects how much liquid is in it," says CSAIL PhD student Andrew Owens, who was lead author on an upcoming paper describing the work. "An algorithm that simulates such sounds can reveal key information about objects' shapes and material types, as well as the force and motion of their interactions with the world."

The team used techniques from the field of "deep learning," which involves teaching computers to sift through huge amounts of data to find patterns on their own. Deep learning approaches are especially useful because they free computer scientists from having to hand-design algorithms and supervise their progress.

The paper's co-authors include recent PhD graduate Phillip Isola and MIT professors Edward Adelson, Bill Freeman, Josh McDermott, and Antonio Torralba. The paper will be presented later this month at the annual conference on Computer Vision and Pattern Recognition (CVPR) in Las Vegas.

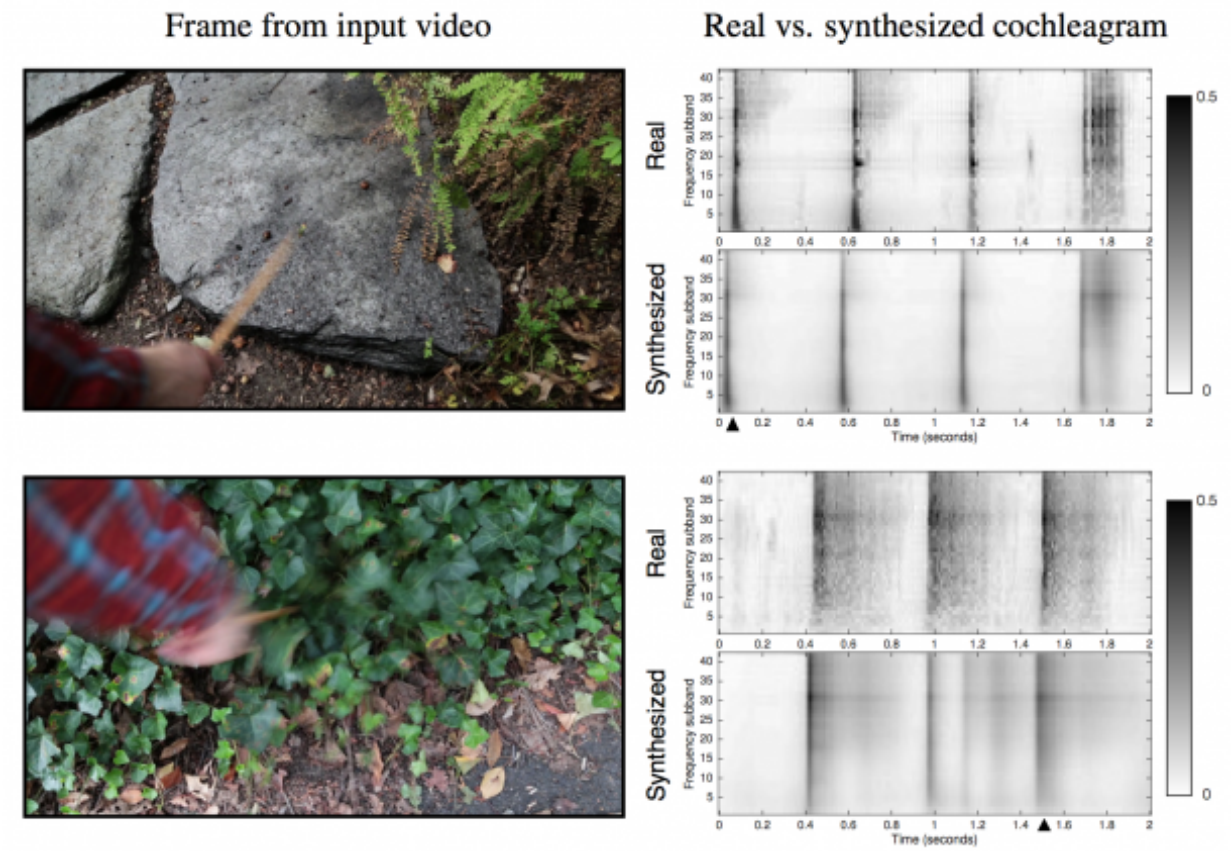
How it works

The first step to training a sound-producing algorithm is to give it sounds to study. Over several months, the researchers recorded roughly 1,000 videos of an estimated 46,000 sounds that represent various objects being hit, scraped, and prodded with a drumstick. (They used a drumstick because it provided a consistent way to produce a sound.)

Next, the team fed those videos to a deep-learning algorithm that deconstructed the sounds and analyzed their pitch, loudness and other features.

"To then predict the sound of a new video, the algorithm looks at the sound properties of each frame of that video, and matches them to the most similar sounds in the database," says Owens. "Once the system has those bits of audio, it stitches them together to create one coherent sound."

The result is that the algorithm can accurately simulate the subtleties of different hits, from the staccato taps of a rock to the longer waveforms of rustling ivy. Pitch is no problem either, as it can synthesize hit-sounds ranging from the low-pitched "thuds" of a soft couch to the high-pitched "clicks" of a hard wood railing.



Credit: Massachusetts Institute of Technology

"Current approaches in AI only focus on one of the five sense modalities, with vision researchers using images, speech researchers using audio, and so on," says Abhinav Gupta, an assistant professor of robotics at Carnegie Mellon University who was not involved in the study. "This paper is a step in the right direction to mimic learning the way humans do, by integrating sound and sight."

An additional benefit of the work is that the team's library of 46,000 sounds is free and available for other researchers to use. The name of the dataset: "Greatest Hits."

Fooling humans

To test how realistic the fake sounds were, the team conducted an online study in which subjects saw two videos of collisions—one with the actual recorded sound, and one with the algorithm's—and were asked which one was real.

The result: Subjects picked the fake sound over the real one twice as often as a baseline algorithm. They were particularly fooled by materials like leaves and dirt that tend to have less "clean" sounds than, say, wood or metal.

On top of that, the team found that the materials' sounds revealed key aspects of their physical properties: An algorithm they developed could tell the difference between hard and soft materials 67 percent of the time.

The team's work aligns with recent CSAIL research on audio and video amplification. Freeman has helped develop algorithms that amplify movements captured by video that are invisible to the naked eye, which has allowed his groups to do things like make the human pulse visible and even recover speech using nothing more than video of a potato chip bag.

Looking ahead

Researchers say that there's still room to improve the system. For example, if the drumstick moves especially erratically in a video, the algorithm is more likely to miss or hallucinate a false hit. It is also limited by the fact that it applies only to "visually indicated sounds"—sounds that are directly caused by the physical interaction that is being depicted in the video.

"From the gentle blowing of the wind to the buzzing of laptops, at any given moment there are so many ambient sounds that aren't related to what we're actually looking at," says Owens. "What would be really exciting is to somehow simulate sound that is less directly associated to the visuals."

The team believe that future work in this area could improve robots' abilities to interact with their surroundings.

"A robot could look at a sidewalk and instinctively know that the cement is hard and the grass is soft, and therefore know what would happen if they stepped on either of them," says Owens. "Being able to predict sound is an important first step toward being able to predict the consequences of physical interactions with the world."

More information: Visually Indicated Sounds. arxiv.org/abs/1512.08512

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