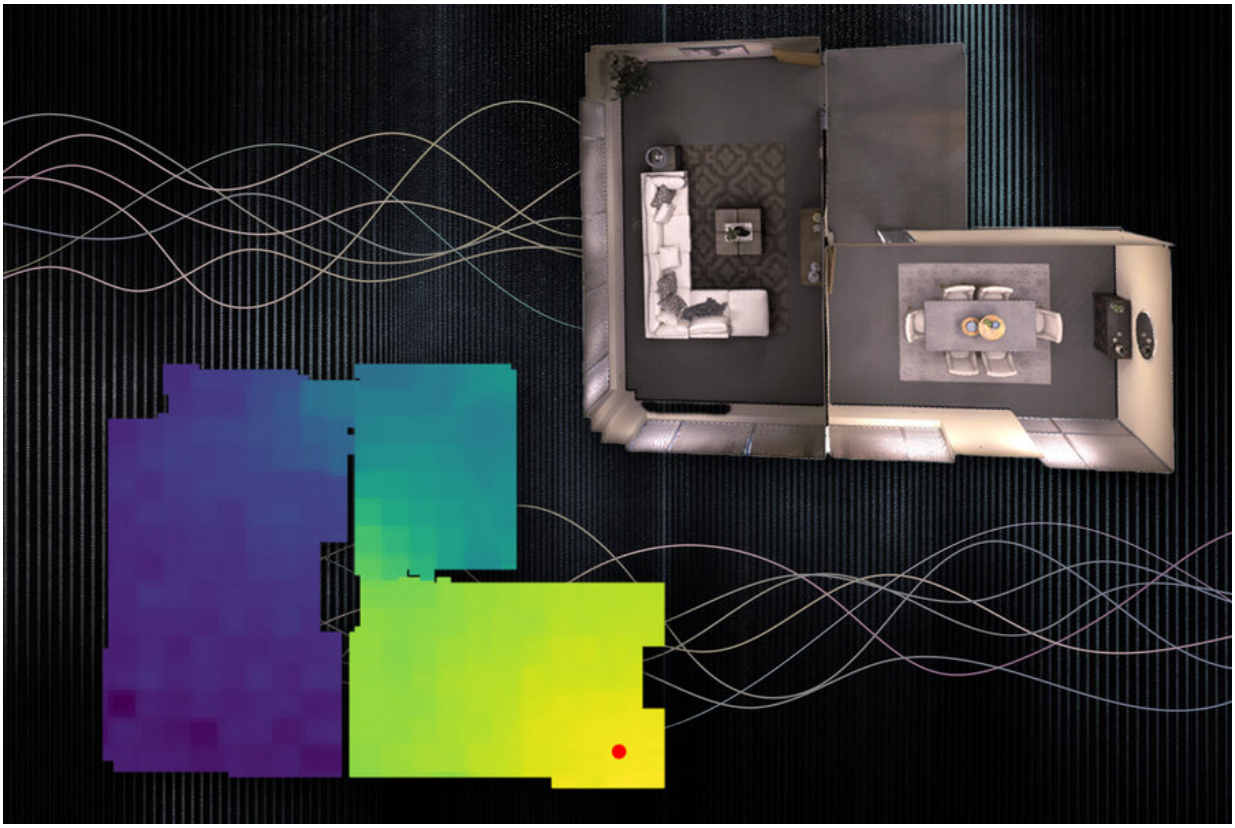
Using sound to model the world

November 1 2022, by Adam Zewe



MIT researchers have developed a machine-learning technique that accurately captures and models the underlying acoustics of a scene from only a limited number of sound recordings. In this image, a sound emitter is marked by a red dot. The colors show the sound volume if a listener were to stand at different locations — yellow is louder and blue is quieter. Credit: Massachusetts Institute of Technology

Imagine the booming chords from a pipe organ echoing through the cavernous sanctuary of a massive, stone cathedral.

The sound a cathedral-goer will hear is affected by many factors, including the location of the organ, where the listener is standing, whether any columns, pews, or other obstacles stand between them, what the walls are made of, the locations of windows or doorways, etc. Hearing a sound can help someone envision their environment.

Researchers at MIT and the MIT-IBM Watson AI Lab are exploring the use of spatial acoustic information to help machines better envision their environments, too. They developed a machine-learning model that can capture how any sound in a room will propagate through the space, enabling the model to simulate what a listener would hear at different locations.

By accurately modeling the acoustics of a scene, the system can learn the underlying 3D geometry of a room from sound recordings. The researchers can use the acoustic information their system captures to build accurate visual renderings of a room, similarly to how humans use sound when estimating the properties of their physical environment.

In addition to its potential applications in virtual and augmented reality, this technique could help artificial-intelligence agents develop better understandings of the world around them. For instance, by modeling the acoustic properties of the sound in its environment, an underwater exploration robot could sense things that are farther away than it could with vision alone, says Yilun Du, a grad student in the Department of Electrical Engineering and Computer Science (EECS) and co-author of a paper describing the model.

"Most researchers have only focused on modeling vision so far. But as humans, we have multimodal perception. Not only is vision important,

sound is also important. I think this work opens up an exciting research direction on better utilizing sound to model the world," Du says.

Joining Du on the paper are lead author Andrew Luo, a grad student at Carnegie Mellon University (CMU); Michael J. Tarr, the Kavčić-Moura Professor of Cognitive and Brain Science at CMU; and senior authors Joshua B. Tenenbaum, the Paul E. Newton Career Development Professor of Cognitive Science and Computation in MIT's Department of Brain and Cognitive Sciences and a member of the Computer Science and Artificial Intelligence Laboratory (CSAIL); Antonio Torralba, the Delta Electronics Professor of Electrical Engineering and Computer Science and a member of CSAIL; and Chuang Gan, a principal research staff member at the MIT-IBM Watson AI Lab. The research will be presented at the Conference on Neural Information Processing Systems.

Sound and vision

In computer vision research, a type of machine-learning model called an implicit neural representation model has been used to generate smooth, continuous reconstructions of 3D scenes from images. These models utilize neural networks, which contain layers of interconnected nodes, or neurons, that process data to complete a task.

The MIT researchers employed the same type of model to capture how sound travels continuously through a scene.

But they found that vision models benefit from a property known as photometric consistency which does not apply to sound. If one looks at the same object from two different locations, the object looks roughly the same. But with sound, change locations and the sound one hears could be completely different due to obstacles, distance, etc. This makes predicting audio very difficult.

The researchers overcame this problem by incorporating two properties of acoustics into their model: the reciprocal nature of sound and the influence of local geometric features.

Sound is reciprocal, which means that if the source of a sound and a listener swap positions, what the person hears is unchanged. Additionally, what one hears in a particular area is heavily influenced by local features, such as an obstacle between the listener and the source of the sound.

To incorporate these two factors into their model, called a neural acoustic field (NAF), they augment the neural network with a grid that captures objects and architectural features in the scene, like doorways or walls. The model randomly samples points on that grid to learn the features at specific locations.

"If you imagine standing near a doorway, what most strongly affects what you hear is the presence of that doorway, not necessarily geometric features far away from you on the other side of the room. We found this information enables better generalization than a simple fully connected network," Luo says.

From predicting sounds to visualizing scenes

Researchers can feed the NAF visual information about a scene and a few spectrograms that show what a piece of audio would sound like when the emitter and listener are located at target locations around the room. Then the model predicts what that audio would sound like if the listener moves to any point in the scene.

The NAF outputs an impulse response, which captures how a sound should change as it propagates through the scene. The researchers then apply this impulse response to different sounds to hear how those sounds

should change as a person walks through a room.

For instance, if a song is playing from a speaker in the center of a room, their model would show how that sound gets louder as a person approaches the speaker and then becomes muffled as they walk out into an adjacent hallway.

When the researchers compared their technique to other methods that model acoustic information, it generated more accurate sound models in every case. And because it learned local geometric information, their model was able to generalize to new locations in a scene much better than other methods.

Moreover, they found that applying the acoustic information their model learns to a computer vison model can lead to a better visual reconstruction of the scene.

"When you only have a sparse set of views, using these acoustic features enables you to capture boundaries more sharply, for instance. And maybe this is because to accurately render the acoustics of a scene, you have to capture the underlying 3D geometry of that scene," Du says.

The researchers plan to continue enhancing the model so it can generalize to brand new scenes. They also want to apply this technique to more complex impulse responses and larger scenes, such as entire buildings or even a town or city.

"This new technique might open up new opportunities to create a multimodal immersive experience in the metaverse application," adds Gan.

"My group has done a lot of work on using machine-learning methods to accelerate acoustic simulation or model the acoustics of real-world

scenes. This paper by Chuang Gan and his co-authors is clearly a major step forward in this direction," says Dinesh Manocha, the Paul Chrisman Iribe Professor of Computer Science and Electrical and Computer Engineering at the University of Maryland, who was not involved with this work.

"In particular, this paper introduces a nice implicit representation that can capture how sound can propagate in real-world scenes by modeling it using a linear time-invariant system. This work can have many applications in AR/VR as well as real-world scene understanding."

More information: Andrew Luo et al, Learning Neural Acoustic Fields, *arXiv* (2022). DOI: 10.48550/arxiv.2204.00628

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