# Alexa and Siri, listen up! Research team is teaching machines to really hear us

July 20 2022

Credit: Pixabay/CC0 Public Domain

University of Virginia cognitive scientist Per Sederberg has a fun experiment you can try at home. Take out your smartphone and, using a voice assistant such as the one for Google's search engine, say the word

"octopus" as slowly as you can.

Your device will struggle to reiterate what you just said. It might supply a nonsensical response, or it might give you something close but still off—like "toe pus." Gross!

The point is, Sederberg said, when it comes to receiving auditory signals like humans and other animals do—despite all of the computing power dedicated to the task by such heavyweights as Google, Deep Mind, IBM and Microsoft—current artificial intelligence remains a bit hard of hearing.

The outcomes can range from comical and mildly frustrating to downright alienating for those who have speech problems.

But using recent breakthroughs in neuroscience as a model, UVA collaborative research has made it possible to convert existing AI neural networks into technology that can truly hear us, no matter at what pace we speak.

The deep learning tool is called SITHCon, and by generalizing input, it can understand words spoken at different speeds than a network was trained on.

This new ability won't just change the end-user's experience; it has the potential to alter how artificial neural networks "think"—allowing them to process information more efficiently. And that could change everything in an industry constantly looking to boost processing capability, minimize data storage and reduce AI's massive carbon footprint.

Sederberg, an associate professor of psychology who serves as the director of the Cognitive Science Program at UVA, collaborated with

graduate student Brandon Jacques to program a working demo of the technology, in association with researchers at Boston University and Indiana University.

"We've demonstrated that we can decode speech, in particular scaled speech, better than any model we know of," said Jacques, who is first author on the paper.

Sederberg added, "We kind of view ourselves as a ragtag band of misfits. We solved this problem that the big crews at Google and Deep Mind and Apple didn't."

The research was presented Tuesday at the high-profile International Conference on Machine Learning, or ICML, in Baltimore.

## Current AI training: Auditory overload

For decades, but more so in the last 20 years, companies have built complex artificial neural networks into machines to try to mimic how the human brain recognizes a changing world. These programs don't just facilitate basic information retrieval and consumerism; they also specialize to predict the stock market, diagnose medical conditions and surveil for national security threats, among many other applications.

"At its core, we are trying to detect meaningful patterns in the world around us," Sederberg said. "Those patterns will help us make decisions on how to behave and how to align ourselves with our environment, so we can get as many rewards as possible."

Programmers used the brain as their initial inspiration for the technology, thus the name "neural networks."

"Early AI researchers took the basic properties of neurons and how

they're connected to one another and recreated those with computer code ," Sederberg said.

For complex problems like teaching machines to "hear" language, however, programmers unwittingly took a different path than how the brain actually works, he said. They failed to pivot based on developments in the understanding of neuroscience.

"The way these large companies deal with the problem is to throw computational resources at it," the professor explained. "So they make the neural networks bigger. A field that was originally inspired by the brain has turned into an engineering problem."

Essentially, programmers input a multitude of different voices using different words at different speeds and train the large networks through a process called back propagation. The programmers know the responses they want to achieve, so they keep feeding the continuously refined information back in a loop. The AI then begins to give appropriate weight to aspects of the input that will result in accurate responses. The sounds become usable characters of text.

"You do this many millions of times," Sederberg said.

While the training data sets that serve as the inputs have improved, as have computational speeds, the process is still less than ideal as programmers add more layers to detect greater nuances and complexity—so-called "deep" or "convolutional" learning.

More than 7,000 languages are spoken in the world today. Variations arise with accents and dialects, deeper or higher voices—and of course faster or slower speech. As competitors create better products, at every step, a computer has to process the information.

That has real-world consequences for the environment. In 2019, a study found that the carbon dioxide emissions from the energy required in the training of a single large deep-learning model equated to the lifetime footprint of five cars.

Three years later, the data sets and neural networks have only continued to grow.

## How the brain really hears speech

The late Howard Eichenbaum of Boston University coined the term "time cells," the phenomenon upon which this new AI research is constructed. Neuroscientists studying time cells in mice, and then humans, demonstrated that there are spikes in neural activity when the brain interprets time-based input, such as sound. Residing in the hippocampus and other parts of the brain, these individual neurons capture specific intervals—data points that the brain reviews and interprets in relationship. The cells reside alongside so-called "place cells" that help us form mental maps.

Time cells help the brain create a unified understanding of sound, no matter how fast or slow the information arrives.

"If I say 'ooooooooc-toooooo-pussssssss,' you've probably never heard someone say 'octopus' at that speed before, and yet you can understand it because the way your brain is processing that information is called 'scale invariant,'" Sederberg said. "What it basically means is if you've heard that and learned to decode that information at one scale, if that information now comes in a little faster or a little slower, or even a lot slower, you'll still get it."

The main exception to the rule, he said, is information that comes in hyper-fast. That data will not always translate. "You lose bits of

information," he said.

Cognitive researcher Marc Howard's lab at Boston University continues to build on the time cell discovery. A collaborator with Sederberg for over 20 years, Howard studies how human beings understand the events of their lives. He then converts that understanding to math.

Howard's equation describing auditory memory involves a timeline. The timeline is built using time cells firing in sequence. Critically, the equation predict that the timeline blurs—and in a particular way—as sound moves toward the past. That's because the brain's memory of an event grows less precise with time.

"So there's a specific pattern of firing that codes for what happened for a specific time in the past, and information gets fuzzier and fuzzier the farther in the past it goes," Sederberg said. "The cool thing is Marc and a post-doc going through Marc's lab figured out mathematically how this should look. Then neuroscientists started finding evidence for it in the brain."

Time adds context to sounds, and that's part of what gives what's spoken to us meaning. Howard said the math neatly boils down.

"Time cells in the brain seem to obey that equation," Howard said.

## UVA codes the voice decoder

About five years ago, Sederberg and Howard identified that the AI field could benefit from such representations inspired by the brain. Working with Howard's lab and in consultation with Zoran Tiganj and colleagues at Indiana University, Sederberg's Computational Memory Lab began building and testing models.

Jacques made the big breakthrough about three years ago that helped him do the coding for the resulting proof of concept. The algorithm features a form of compression that can be unpacked as needed—much the way a zip file on a computer works to compress and store large-size files. The machine only stores the "memory" of a sound at a resolution that will be useful later, saving storage space.

"Because the information is logarithmically compressed, it doesn't completely change the pattern when the input is scaled, it just shifts over," Sederberg said.

The AI training for SITHCon was compared to a pre-existing resource available free to researchers called a "temporal convolutional network." The goal was to convert the network from one trained only to hear at specific speeds.

The process started with a basic language—Morse code, which uses long and short bursts of sound to represent dots and dashes—and progressed to an open-source set of English speakers saying the numbers 1 through 9 for the input.

In the end, no further training was needed. Once the AI recognized the communication at one speed, it couldn't be fooled if a speaker strung out the words.

"We showed that SITHCon could generalize to speech scaled up or down in speed, whereas other models failed to decode information at speeds they didn't see at training," Jacques said.

Now UVA has decided to make its code available for free, in order to advance the knowledge. The team says the information should adapt for any neural network that translates voice.

"We're going to publish and release all the code because we believe in open science," Sederberg said. "The hope is that companies will see this, get really excited and say they would like to fund our continuing work. We've tapped into a fundamental way the brain processes information, combining power and efficiency, and we've only scratched the surface of what these AI models can do."

But knowing that they've built a better mousetrap, are the researchers worried at all about how the new technology might be used?

Sederberg said he's optimistic that AI that hears better will be approached ethically, as all technology should be in theory.

"Right now, these companies have been running into computational bottlenecks while trying to build more powerful and useful tools," he said. "You have to hope the positives outweigh the negatives. If you can offload more of your thought processes to computers, it will make us a more productive world, for better or for worse."

Jacques, a new father, said, "It's exciting to think our work may be giving birth to a new direction in AI."

**More information:** Abstract: proceedings.mlr.press/v162/jacques22a.html

Provided by University of Virginia