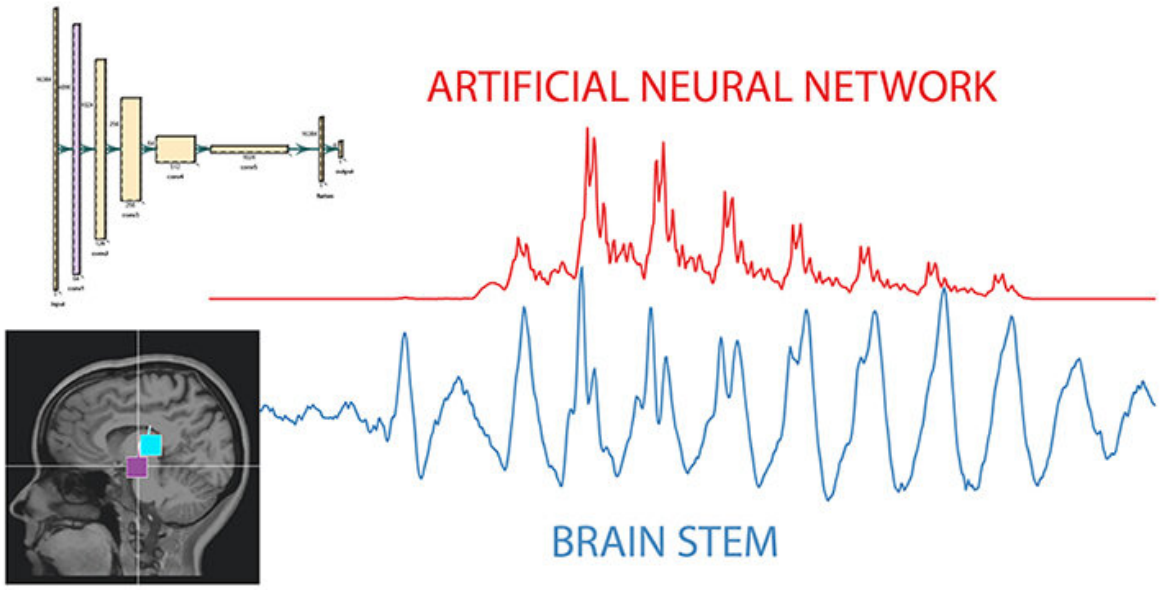
'Raw' data show AI signals mirror how the brain listens and learns

May 2 2023, by Jason Pohl



Researchers found strikingly similar signals between the brain and artificial neural networks. The blue line is brain wave when humans listen to a vowel. Red is the artificial neural network's response to the exact same vowel. The two signals are raw, meaning no transformations were needed. Credit: Gasper Begus

New research from the University of California, Berkeley, shows that artificial intelligence (AI) systems can process signals in a way that is

remarkably similar to how the brain interprets speech, a finding scientists say might help explain the black box of how AI systems operate.

Using a system of electrodes placed on participants' heads, scientists with the Berkeley Speech and Computation Lab measured brain waves as participants listened to a single syllable—"bah." They then compared that brain activity to the signals produced by an AI system trained to learn English.

"The shapes are remarkably similar," said Gasper Begus, assistant professor of linguistics at UC Berkeley and lead author on the study published recently in the journal *Scientific Reports*. "That tells you similar things get encoded, that processing is similar."

A side-by-side comparison graph of the two signals shows that similarity strikingly.

"There are no tweaks to the data," Begus added. "This is raw."

AI systems have recently advanced by leaps and bounds. Since ChatGPT ricocheted around the world last year, these tools have been forecast to upend sectors of society and revolutionize how millions of people work. But despite these impressive advances, scientists have had a limited understanding of how exactly the tools they created operate between input and output.

A question and answer in ChatGPT has been the benchmark to measure an AI system's intelligence and biases. But what happens between those steps has been something of a black box. Knowing how and why these systems provide the information they do—how they learn—becomes essential as they become ingrained in daily life in fields spanning health care to education.

Begus and his co-authors, Alan Zhou of Johns Hopkins University and T. Christina Zhao of the University of Washington, are among a cadre of scientists working to crack open that box.

To do so, Begus turned to his training in linguistics.

When we listen to spoken words, Begus said, the sound enters our ears and is converted into electrical signals. Those signals then travel through the brainstem and to the outer parts of our brain. With the electrode experiment, researchers traced that path in response to 3,000 repetitions of a single sound and found that the brain waves for speech closely followed the actual sounds of language.

The researchers transmitted the same recording of the "bah" sound through an unsupervised neural network—an AI system—that could interpret sound. Using a technique developed in the Berkeley Speech and Computation Lab, they measured the coinciding waves and documented them as they occurred.

Previous research required extra steps to compare waves from the brain and machines. Studying the waves in their raw form will help researchers understand and improve how these systems learn and increasingly come to mirror human cognition, Begus said.

"I'm really interested as a scientist in the interpretability of these models," Begus said. "They are so powerful. Everyone is talking about them. And everyone is using them. But much less is being done to try to understand them."

Begus believes that what happens between input and output doesn't have to remain a black box. Understanding how those signals compare to the brain activity of human beings is an important benchmark in the race to build increasingly powerful systems. So is knowing what's going on

under the hood.

For example, having that understanding could help put guardrails on increasingly powerful AI models. It could also improve our understanding of how errors and bias are baked into the learning processes.

Begus said he and his colleagues are collaborating with other researchers using brain imaging techniques to measure how these signals might compare. They're also studying how other languages, like Mandarin, are decoded in the brain differently and what that might indicate about knowledge.

Many models are trained on visual cues, like colors or written text—both of which have thousands of variations at the granular level. Language, however, opens the door for a more solid understanding, Begus said.

The English language, for example, has just a few dozen sounds.

"If you want to understand these models, you have to start with simple things. And speech is way easier to understand," Begus said. "I am very hopeful that speech is the thing that will help us understand how these models are learning."

In cognitive science, one of the primary goals is to build mathematical models that resemble humans as closely as possible. The newly documented similarities in brain waves and AI waves are a benchmark on how close researchers are to meeting that goal.

"I'm not saying that we need to build things like humans," Begus said. "I'm not saying that we don't. But understanding how different architectures are similar or different from humans is important."

More information: Gašper Beguš et al, Encoding of speech in convolutional layers and the brain stem based on language experience, *Scientific Reports* (2023). DOI: 10.1038/s41598-023-33384-9

Provided by University of California - Berkeley