Deciphering the Speech Code:

Giving a Voice to Paralyzed Patients

INTERVIEW WITH: Dr. David Moses

David Moses, PhD, is an assistant professor of neurological surgery at the University of California, San Francisco. He specializes in bioengineering and biomedical engineering. Currently, his research focuses on speech decoding and the implementation of AI in neuroprosthetics. In the past, he has released publications on machine translation of cortical activity to text, neuroprostheses for decoding speech, neural speech recognition, and artificial intelligence in the communication sciences and otolaryngology, among other related topics.

BY: Aneesa Mustafa, Carlyn Leavitt, Sania Choudhary, and Tanya Sanghal

BSJ: What motivated you to focus specifically on speech decoding and AI within the realm of neuro-prosthetics? Are there personal or professional experiences that influenced your decision?

DM: For me, I knew when I graduated undergrad that I wanted to work on something related to the brain. I started out as a computer science major for the first week of undergrad, then quickly switched to bioengineering because I felt like I would be able to apply computer science to be able to help someone someday. One of the main reasons I joined the UC Berkeley–UCSF Joint Program in bioengineering was because there were so many labs that do brain–computer interface neuroprosthesis research. When I got here, I did not know that I wanted to do speech, specifically. When I heard Dr. Edward Chang, my PI, speak at a seminar class, I learned about his work with speech—how the brain processes speech, how they can use things from natural language processing and machine learning to inform what we can learn about the brain, and how you can decode information for brain signals related to that. I thought it was super awesome. I rotated in the lab, and 10 years later, I am still here. I did my PhD, post-doc, and now I am an adjunct professor in his lab. The journey has been learning about how we can decode speech information from the brain, an opportunity to do a clinical trial, and the feasibility of translating that technology to patients who would need it. Through this journey, I have gained more passion for it, especially working with patients. It is really humbling to work with these incredible people who volunteer so much of their time to work with us. For me, it was really profound that they have so much that they want to express, but it is so difficult. I have worked with a patient who, for over an hour, was trying to use an eye tracker, but it kept messing up and he could not get a sentence out. Seeing that firsthand really made me think we can do better. That is really what I am most passionate about now: how can we take this technology and use it to help someone reconnect with the world?

"For me, it was really profound that they have so much that they want to express, but it is so difficult."

BSJ: Could you explain the process of creating the devices/codes used in the studies?

DM: We get our hardware from manufacturers who make the array, which is connected to a pedestal. The array sits on the surface of patient Anne's brain, and the pedestal is actually implanted, so it goes through her skull. Then we connect a connector to the external portion of this pedestal, which allows us to acquire the signals from the array. The signals go through our whole system, are processed some more, and then get through our code that we run our algorithms on to do the decoding of speech. We focus mostly on the algorithms, which are a combination of things that started out when I was working on my PhD, but have now evolved way beyond that.

The array is implanted over the speech motor cortex, the brain area that controls our vocal tract and speech. This patient, Anne, who we worked with for the recent study, had a brainstem stroke. The signals that would have normally controlled the vocal tract to enable speech, like how we speak, are severed at the brainstem and can not reach her facial muscles. So, we implant our centers over that brain area so we can bypass the paralysis. For those commands that would have been sent to a vocal tract, we interpret them and translate them into the output. To do that, we first train the algorithms using deep-learning models. When Anne first sees a sentence, there will be a brief countdown before

the sentence turns green, and when it does, she attempts to "mime" the sentence—she tries to speak but silently. During that time, we collected brain activity. We had three different outputs: text, speech, and then an avatar.

For the text output, one thing that we do is extract what we call phoneme probabilities. Think of this like a heat map, where one dimension is time, and one dimension is phonemes. Phonemes are like the alphabet of spoken language—the different sounds that, when you change one sound, you change the meaning of a word. For example, "cat" has three phonemes. There is a C, an A, and a T. If I were to change the short vowel A to the broad sound "aw," then all of a sudden, I would have "cawt" (cot). Phonemes have been used for speech recognition. When you are talking to your phone, there is a chance that it actually goes through phonemes and then into text to figure out what you are trying to say.

The beauty of this approach is that we can then use things from natural language processing trained on millions of hours of speech or text to do the rest. We apply what we call a lexical constraint, or the likely sentences that are associated with this phoneme probability. Essentially, this generates a bunch of hypotheses and scores to ask how likely it was. The final step is to apply a language model using information about word sequences in English. For example, the question, "how are you?" is much more likely to appear than "how are glasses?" This step involves applying vocabulary constraints, turning these phoneme probabilities into words, and rescoring sentences to say what is likely given the structure of English. That is how we get our final output.

For the speech portion, instead of mapping the phonemes, we map to something we call 'discrete speech units'. We take a sound waveform, and convert it to a different representation that simplifies the speech. We can do some extra steps, like acoustic processing, to generate the actual speech waveform in something that is personalized to the participant's voice. Anne shared with us some footage of her before her injury, and we took that to model and generate a voice profile that matches her likeness.

For articulators, instead of speech units, we have these articulator units that we can use to animate an AI face. We work with a company called Speech Graphics, and they go from acoustics to these articulatory representations. They do this for video games or shows to animate an avatar—maybe you get a voice actor that does a voice, so you can animate the character. In real-time, we can decode the speech and use that to drive the avatar in perfect sync. This way, we were able to get all three outputs at the same time—the text, the speech, and the avatar, all originating from the same brain activity.

BSJ: What was the process of getting the participant familiar and interacting with the system? Were there any challenges the team faced regarding either the participant or the technology? How did you address them?

DM: With Anne's everyday interface, she is used to it. It is based on face and head tracking—she wears glasses that have a dot, and the system can tell where she is looking. But it is not very fast, it is about 10 times slower than when we can speak to each other. This affects, categorically, what types of interactions you can have. She and her husband were describing that if they were getting into an argument or a really intense discussion, that cadence is completely ruined when they had to wait three to four minutes for her to respond. He was describing that he knows it is not fair, but he becomes a little bit detached. In other words, you say something at the moment, then you have to wait for them to respond, so you have the time to think about something else—go get a cup of water or something, and then you come back; it is so unnatural. In other words, Anne is used to her interface, but it does make it really difficult to have some types of interactions.

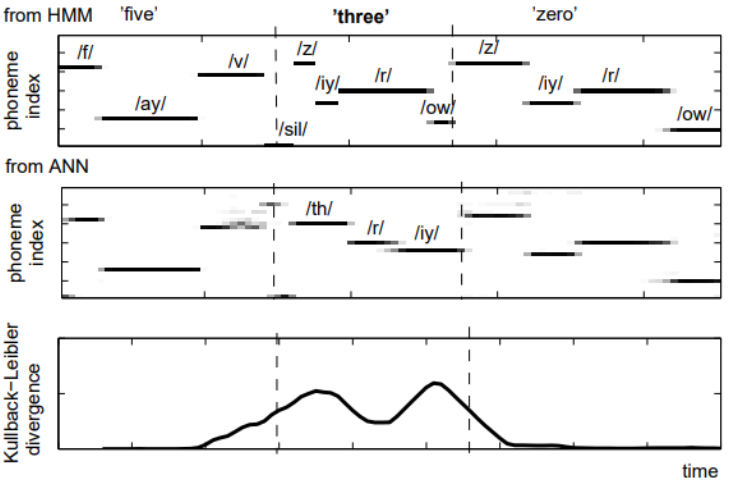

Figure 1. Probabilities of estimated phonemes. This graph represents the utterance of "five three zero." The word "three" shows that it's an unexpected word that is not in the vocabulary.

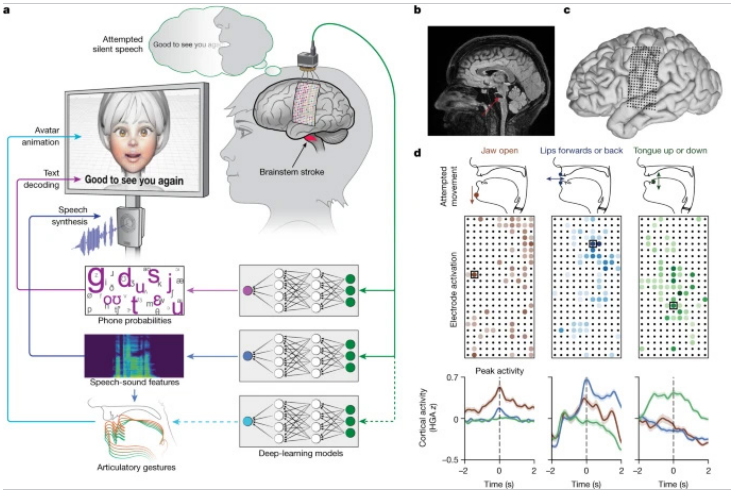

Figure 2: Multimodal speech decoding. This schematic is an overview of the speech-decoding pipeline. **a**. Neural activity is processed and deep-learning models are applied. **b**. MRI scan showing brainstem atrophy after stroke. **c.** MRI reconstruction showing the position of the implanted electrodes. **d**. Basic articulatory movements, electrode-activation maps indicating robust accuracy across articulators during attempts at articulatory movements, and cortical activity responses with each movement type.

"We do hope that over time, something very natural and intuitive can be made—a system that can enable some of the users in their daily lives."

For our system, though, Anne cannot use it in her daily life. This is something that we are working towards. In our clinical trial, the pedestal is very sensitive and fragile. If someone were to make a mistake and scratch it, the participant might have to get another surgery to fix it. There are also all kinds of maintenance that we have to do to the area around the implant to make sure that it does not get infected. We do hope that over time, something very natural and intuitive can be made—a system that can enable some of the users in their daily lives. Regarding getting familiar with our system, Anne found that the instructions were pretty straightforward. She was able to pick up on what she needed to do fairly quickly. We just had to get a lot of training data from her for about two weeks.

BSJ: The studies have only been conducted with adult participants. What specific goals need to be reached before speech decoding can be done with younger paralyzed people?

DM: It's something that we definitely think about, because there's a lot of need there. Not just for stroke, paralysis, or even other trauma/atrophy conditions in certain brain regions, but also for conditions like nonverbal Autism Spectrum Disorder (ASD) That being said, there are a lot of challenges. For someone who's younger, their brain is probably still growing. So, how do you implant an array so that as the brain grows, it doesn't get messed up? And then I think for other conditions like ASD, it's complicated because it is such a heterogeneous disorder, that it is difficult because there might be some

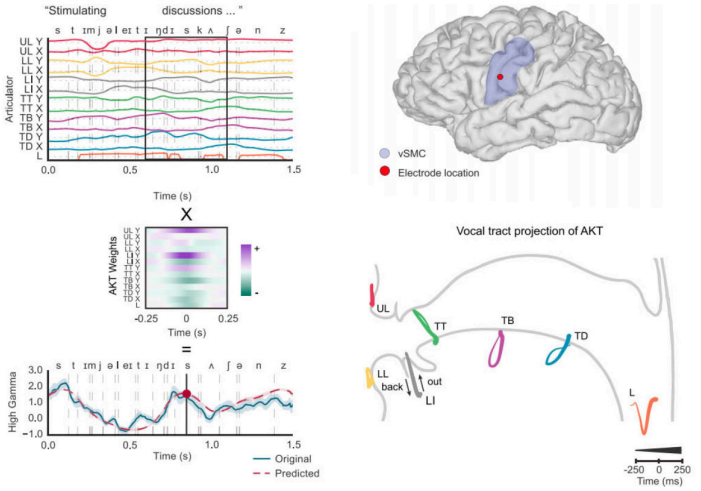

Figure 3: INSERT TITLE. The Chang Lab found that the jaw, tongue, lips, and throat have coordinated movement at single electrodes where articulatory pathways were encoded for by neural populations. Over many studies, they discovered the diversity of such encoded movements, which represent various speech sounds in English.

who don't understand language like we understand it; for our system to work they have to try to speak – with what we've validated, if someone isn't able to try to say the words, then it wouldn't work. And, as the brain is changing, maybe the signals change over time, and these types of things where it is a very dynamic system as the source, it is just a little bit more complicated.

BSJ: How did you control the timing alignment between intended speech and neural activity during your process of model training?

DM: We have a separate algorithm that detects speech from the brain activity. When Poncho, our first participant, was attempting to say something, we could detect it from his neural signature. That is what we used to get our alignment. It was most important during testing, because it was crucial that he could speak at his own rate.

BSJ: In the methods, it is mentioned that sometime during the training, the participants' sentences with syntactic pauses or punctuation in the middle were removed from the 1024-word general sentence. How does punctuation pose as a limitation in this pursuit?

DM: We wanted, for simplicity of training, the rate at which Anne produces words to be roughly constant. We did not want pauses in the sentences, because it might be a little bit confusing for the algorithm. We don't really fully understand where other parts of syntax, like commas, are represented in the brain. Although I personally do not think they're represented very strongly in the brain area that we record from, we wanted it to be simpler so that the model can really focus on what we want, which is to learn this mapping between brain activity and phonemes, etc. Now, there are new language models which can handle those things very well, so this is not, I would say, an impossibility. It was only for this proof of concept that we decided not to do that.

BSJ: This paper stood out because it included an avatar that took the task for decoding and gave it expressions. Can you explain more regarding the process and reasoning of adding this feature of facial expressions to create a more immersive communication experience?

DM: Throughout the course of the project conception, when we originally were thinking about doing the avatar, we were thinking of using it as a tool to help train the participants. We thought that it could be really informative in showing what we want reading a sentence to look like, and the rate, and here's what you could try to use it to embody yourself as like, this is what my target is to move my mouth in this way to speed this thing. Now, obviously, it was very different by the end. And I think a lot of that came down to well, we have this opportunity to work with this company, Speech Graphics, who has done some amazing work on the engineering side, on the animation side, and we can do something that's, you know, the world is getting more and more digital. And things like an avatar, I mean, even the movies, right, Avatar movies, it's like people are interested in this. But the metaverse I mean, these are all things that are kind of like they may be part of our future. I don't know to what extent but likely in 1020 years, it will be a little more widespread. And as people live these more digital lives, being able to have someone embodied in a virtual space. I think is is a good goal. And it felt reasonable to add in given the flexibility of the modeling that we were able to see where we're going to go text and speech and let's add an avatar to add this next level of embodiment there. I mean, one thing that I think about, I know that some people when they hear Metaverse, they think of Facebook Meta, like Metaverse, and like, what is this? You know, I don't want this. What's the point? And I think those are all fair opinions. But what I think about in this context is a lot of the reasons why maybe a

Metaverse or things like this don't seem that exciting to many people. It's like, Why do I want to go in VR, and walk into a boardroom and have a meeting or like do something like this, like, I can just go do that. But actually, it's very equalizing, if you can have someone with very severe physical disabilities, like they might actually be able to do more in VR, physical like in an interactive way than they could do in real life. And I think that, that opportunity, like the immersiveness, and the kind of almost restorative aesthetic could have and just like the free, right word, it's almost like being able to free someone in some way. It is pretty interesting. And I like that's far, probably down the line a little bit. But I think that those types of things are very interesting for us to maybe think about now is how a digital virtual interaction can, can be done. Regardless of physical disability. And so that's interesting. And I'll add one last tidbit here, which is that and some of her feedback to us, I didn't mention, but she actually chose the avatar, this approach can work with any of a wide variety of avatars, and we can customize them. And it uses an Unreal Engine. And meta humans just like this code base that we use, and it's very flexible, so we can have many different avatars. And she actually chose the one that you see in the paper out of a list of many, many candidates. And so he chose him because she thought it was because she liked having it to embody herself. She thought it was very cool. And she told us that one of her dreams is to become a counselor for someone, for people who have gone through this similar

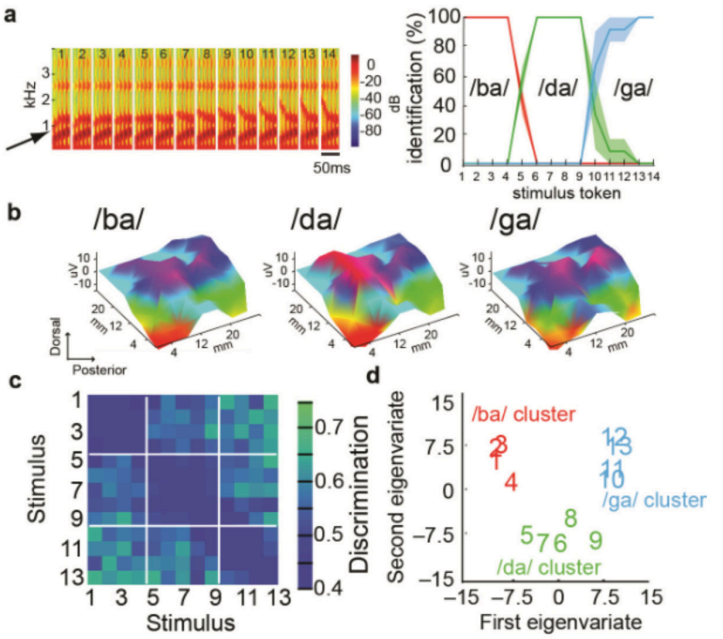

Figure 4: This image shows how phonemes (building blocks of spoken language) are read from an acoustic stream. Specifically, in this case, the phonemes 'ba', 'da', and 'ga' are being extracted.

experience as what you've gone through, become severely paralyzed. And being able to like work with them and counsel them so that they know, just for their mental health and things like this, so she thought the avatar would be really helpful for that. She thought it could be a really nice, nice way to interact with potential patients lines that will put them at ease. So it was really interesting for us to hear that feedback. So I think it is interesting and we just scratched the surface with it. But I think there's a lot of cool stuff that can be done if you could control a digital likeness of yourself.

BSJ: What future research studies do you have planned?

DM: We want to get the accuracy higher, we are at about 75% accuracy. We are thrilled by that step, but obviously, we need it to be higher for someone to rely on this for their daily life. We also want to expand the vocabulary: how can we get to more words so that they can be more expressive? There are other things that are very interesting to us as well, like controlling pitch or emphasizing certain words to distinguish a regular sentence from a question. That inflection is something that we are trying to restore.

We also want to get our models to be faster and at a lower latency. Not just in terms of words per minute, but in terms of decreasing the delay between thought and output. Ideally, Anne could learn to use it as a new extension of herself.

On the hardware side, there is a lot of interest in improving it to be a real solution for patients. They need something that is fully implantable and wireless so after the surgery, nothing is exposed. The wound will heal, and then they have an implanted device that they can connect wirelessly to control what they need to. And, lastly we're really interested in how patients would actually be able to use it. For example, how can someone send an email with this or write a post. Poncho, one of our patients, likes to do web development sometimes. He will write letter by letter, create web pages from code. And so the question becomes, how can we make a system that will enable him to do that, without having to write letter by letter?

BSJ: What impacts do you see your research having both in terms of technology development and on the lives of patients with neurological disorders?

DM: The dream is to provide some meaningful solutions to these patients who have these kinds of disabilities. It is about restoring autonomy, independence, and restoring connection and also productivity—the patients we work with want to produce, to create things. They want to contribute to society as well, even though they have a really difficult condition. It is being able to restore all of those things in one solution.

BSJ: There is an idea in anthropology called 'linguistic determinism', the theory that our thoughts and culture are dictated by our language. How do you think linguistic determinism plays a role in decoding the intended speech of an individual with vocal paralysis? How will our culture have to change with AI?

DM: For us, we are decoding from the part of the brain that is controlling our vocal tract, which is less abstract. It's really like controlling the dynamics of our facial muscles. And so we don't think that our methods are extraordinarily affected by abstract language representations. But, I do think that in general, there's an increasing use of AI to actually generate language. If you consider the database of everything that's ever been written, as more and more of that becomes generated by AI, it is going to have effects. And I am not sure how it will affect what society really tries to focus on, like how we can use these new tools for good.

REFERENCES

1. Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G. K., & Chang, E. F. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 620(7976), 1037–1046. https://doi.org/10.1038/s41586-023-06443-4.
2. Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., & Chang, E. F. (2021). Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. The New England journal of medicine, 385(3), 217–227. https://doi.org/10.1056/NEJMoa2027540.

IMAGE REFERENCES

1. Figure 1: Ketabdar, Hamed & Hannemann, Mirko & Hermansky, Hynek. (2007). Detection of out-of-vocabulary words in posterior based ASR. 4. 1757-1760. 10.21437/Interspeech.2007-492.
2. Figure 2: Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G. K., & Chang, E. F. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 620(7976), 1037–1046. https://doi.org/10.1038/s41586-023-06443-4.
3. Figure 3: Chartier, J., Anumanchipalli, G.K., Johnson, K., & Chang, E.F. (2018). Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex. Neuron, 98(5): 1042-1054. doi:10.1016/j.neuron.2018.04.031.
4. Figure 4: Chang, E.F., Rieger, J., Johnson, K.D., Berger, M.S., Barbaro, N.M., & Knight, R.T. (2010). Categorical speech representation in the human superior temporal gyrus. Nat Neurosci. 2010 Nov;13(11):1428-32.