Title: Interpreting Speech and Sounds from Neural Activity, a Brief Overview

Andrew Ryan

Abstract:

For people who are mute, or are completely paralyzed, one of the primary problems they have to deal with is being able to communicate. One potential solution to compensate for decreased communication functions is by using a brain-computer interface (BCI). The idea would be to quantify neural activation in the brain that correlates to imagined speech from the patient, and decode that into legible text that can be interpreted by the receiver. Due to the intricacy of speech interpretation, direct access to regions of the brain and individual neurons is required. As a result, many tests done on BCI speech interpretation involve using ECoG sensors on epilepsy patients when they are available. Some approaches used to analyze these signals for feature extraction include word based classification, and phoneme based classification. One approach mentioned less in the literature, is if there is a method to pull a sound signal directly from the activated regions of the brain. Advancement of the technology has potential use as a speech replacement for people suffering from paralysis, as well as in prosthetics.

Introduction

A brain computer interface (BCI) is an interface that allows direct communication from neuronal firings in the brain to a digital computer output that can be interpreted for practical applications. One of the most versatile potential applications is speech recognition. Not only would it allow mute individuals to communicate freely and effectively, the ability to read words directly from the brain without creating sound could in theory be used as the control for many different prosthetics.

When it comes to interpreting neural signals for words and sounds, there are three primary forms of signal production: actual verbalization of the target word [1], miming of the target word, and imagining of the target word [2]. Additionally, researchers tend to look at three different levels of language construction: full sentences, individual words [3], and phonics [4] . One of the most apparent differences in these studies is the use of words versus phonics for sound classification. Levels of success for imagined sounds typically range between 15-25% in terms of accuracy for both phonics and words. While this is good when compared to random chance, it is still not at the level where effective communication and bit rate can be achieved. While words and phonics are barebones breakdowns of what composes speech, one question that can be raised is whether there could be a more fundamental approach to reading imagined thought. Specifically, would it be possible to differentiate sound regions in the brain based on sounds composed of a single frequency? Spoken words and phonics can be broken down into their sine wave decompositions, and it begs the question as to whether this is the case in mental sound construction as well.

It has been shown through functional magnetic resonance imaging (fMRI) imaging that the location of brain activity, in terms of interpreting words, changes based on how familiar a person is with a word as well as whether or not it is a 'pseudo word' or not [5]. This study also shows that there is an entirely separate branch used for phonetic processing, specifically for uncommon words and pseudo words that the patient needs more time to consider. This branch is accessed when the brain is converting read text into the sounds that the word would create. A paper by André Aleman et al, also shows that activated brain regions for both interpreted sound and imagined sounds are handled in the temporal lobes of the brain [6].

The most commonly seen method for BCI's for the purpose of speech recognition, is by using electrocorticography (ECoG). This is because speech requires a high resolution in a specific area of the brain. There have been attempts however, to use EEG as well. However this would reduce the potential accuracy and consistency of word recognition. The biggest disadvantage of using ECoGs over electroencephalogram (EEG) is that only patients already undergoing surgery for epilepsy or similar disorders can be used for preliminary tests, making them rare and hard to repeat.

This paper will be an analysis on the current state of BCI technology as it relates to word and sound recognition. Focus will be on the success and weaknesses of different styles of BCI for speech recognition and current analysis techniques. Questions will also be made on different potential techniques for BCI sound recognition that have not been thoroughly researched in the literature.

Word vs Phenome Classification

There have also been a few different approaches when it comes to speech interpretation. The two most researched, are individual word recognition, and phenome recognition.

A paper by Stephanie Martin et al. does word based classification of speech using ECoGs in epilepsy patients [2]. In their study, they compared classification accuracy of speech production in three different manners: Listening, where the patient is simply listening to the target word; overt speech, where the patient creates the target word; and imagined speech, where the patient must imagine the target word, and never actually hear it. Results showed that imagined speech could be classified with statistically significant accuracy. Accuracies typically ranged from 50% to 80% depending on the subject and number of words classified, which is much better than random, but is still not useful for practical application.

One of the major disadvantages of using word based classification, is that you must have a classifier for each and every word that you would intend someone to use. Even if you were to limit classification to only the most common of words, that is still a massive undertaking that would require significant computing power, and would need to be individualized to each person, for each word. An alternative, and likely more flexible, method for classification is to instead look at how phenomes are constructed in the brain.

Like words form a sentence, phenomes are used to construct words. Humans have the ability to categorize phenomes for sound recognition and speech interpretation for daily use. If we were to take a continuous graph of potential sounds, we would see that this is discretized into different phenome types [7]. We can also see this categorization in the temporal gyrus, which plays a part in speech construction and perception [8].

A 2014 paper by Emily Mugler et al, did just that, and looked at classifying all 24 English phenomes in the International Phonetic Alphabet (IPA) [1]. This differed from the test above, in that the patients were actually vocalizing the phonics by reading aloud presented words. This resulted in an accuracy that was also around 30-40%, however they noted that phenomes were often misclassified as phenomes that neighbored them (such as the ‘t’ and ‘d’ sounds).

A more recent 2019 paper by Janaki Sheath et al, does classification of words by identifying for phenomes in imagined speech [3]. Their classification accuracy was between 30-40%, which is also significantly lower than the individual word accuracy. This is however still better than random, and they did achieve bitrates that were much faster than other systems. One thing would have been interesting, would be to see what the accuracy would have been if they had just looked at phenomes, rather than running them through a language model that could have reduced accuracy.

Reading a Raw Sound Signal from the Brain

One potential technique that could be used to interpret words from brain activity would be to look at the problem in a purely physical sense, rather than a lexical one. The goal would be to analyze a raw sound signal generated by the brain. It has been shown that unique monotone frequencies can be interpretable from each other using EEG signals, with an ANFIS neural network [9]. This method so far however, can only interpret 3 different frequencies.

Tests that would need to be done to show that this would be a viable method for speech interpretations are: 1) determine if monotone sound frequencies can be read from brain activity on a continuous scale, 2) determine if combinations of different tones are linear and time invariant when processed in the brain, 3) determine if imagined tones and combinations of tones can be effectively recreated by the patient in their imagination, and that these imagined sounds can be extracted in a similar manner to the tones in the previous tests. If all of these were to hold true, you could in theory extract a sound wave of imagined phonemes that could then be constructed into words.

One way to determine how feasible this might be is to look at research done on sound processing in the brain.

Conclusion

Many techniques have been tried to improve the results of these types of BCI. Many of these techniques, while they show a glimpse of feasibility, are ultimately not at the level required for anyone to spend the cost to use them meaningfully. There is still a lot of research and work that needs to be done for creating a BCI that can convert imagined words to text. It is likely that a major breakthrough would need to be done before there can be any meaningful process made.

References

[1] E. M. Mugler *et al.*, "Direct classification of all American English phonemes using signals from functional speech motor cortex," *J. Neural Eng.*, vol. 11, no. 3, p. 035015, May 2014, doi: 10.1088/1741-2560/11/3/035015.

[2] S. Martin *et al.*, "Individual Word Classification During Imagined Speech Using Intracranial Recordings," in *Brain-Computer Interface Research: A State-of-the-Art Summary 7*, C. Guger, N. Mrachacz-Kersting, and B. Z. Allison, Eds. Cham: Springer International Publishing, 2019, pp. 83–91.

[3] J. Sheth, A. Tankus, M. Tran, N. Pouratian, I. Fried, and W. Speier, "Translating neural signals to text using a Brain-Machine Interface," *ArXiv190704265 Cs*, Jul. 2019, Accessed: Feb. 08, 2020. [Online]. Available: http://arxiv.org/abs/1907.04265.

[4] J. S. Brumberg, E. J. Wright, D. S. Andreasen, F. H. Guenther, and P. R. Kennedy, "Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex," *Front. Neurosci.*, vol. 5, 2011, doi: 10.3389/fnins.2011.00065.

[5] C. J. Fiebach, A. D. Friederici, K. Müller, and D. Y. von Cramon, "fMRI Evidence for Dual Routes to the Mental Lexicon in Visual Word Recognition," *J. Cogn. Neurosci.*, vol. 14, no. 1, pp. 11–23, Jan. 2002, doi: 10.1162/089892902317205285.

[6] A. Aleman, E. Formisano, H. Koppenhagen, P. Hagoort, E. H. F. de Haan, and R. S. Kahn, "The Functional Neuroanatomy of Metrical Stress Evaluation of Perceived and Imagined Spoken Words," *Cereb. Cortex*, vol. 15, no. 2, pp. 221–228, Feb. 2005, doi: 10.1093/cercor/bhh124.

[7] A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, "The discrimination of speech sounds within and across phoneme boundaries.," *J. Exp. Psychol.*, vol. 54, no. 5, p. 358, 19590201, doi: 10.1037/h0044417.

[8] E. F. Chang, J. W. Rieger, K. Johnson, M. S. Berger, N. M. Barbaro, and R. T. Knight, "Categorical Speech Representation in Human Superior Temporal Gyrus," *Nat. Neurosci.*, vol. 13, no. 11, pp. 1428–1432, Nov. 2010, doi: 10.1038/nn.2641.

[9] R. Sudirman, A. C. Koh, N. M. Safri, W. B. Daud, and N. H. Mahmood, "EEG different frequency sound response identification using neural network and fuzzy techniques," in *2010 6th International Colloquium on Signal Processing its Applications*, May 2010, pp. 1–6, doi: 10.1109/CSPA.2010.5545237.