INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Identifying input features for development of real-time translation of neural signals to text

Janaki Sheth[1]*, Ariel Tankus*[2]*, Michelle Tran*[3]*, Lindy Comstock*[4]*, Itzhak Fried*[3]*, William Speier*[5]

[1]Dept. of Physics and Astronomy, University of California, Los Angeles, USA
[2] Dept. of Neurology and Neurosurgery, Tel Aviv University, Tel Aviv, ISR
[3] Dept. of Neurosurgery, University of California, Los Angeles, USA
[4] Dept. of Linguistics, University of California, Los Angeles, USA
[5]Dept. of Radiological Sciences, University of California, Los Angeles, USA

janaki.sheth@physics.ucla.edu, arielt@post.tau.ac.il, metran.cnl@gmail.com, lbcomstock@ucla.edu, IFried@mednet.ucla.edu, Speier@ucla.edu

Abstract

One of the main goals in Brain-Computer Interface (BCI) research is to help patients with faltering communication abilities due to neurodegenrative diseases produce text or speech output using their neural recordings. However, practical implementation of such a system has proven difficult due to limitations in the speed, accuracy, and training time of existing interfaces. In this paper, we contribute to this endeavour by isolating appropriate input features from speech-producing neural signals that will feed into a machine learning classifier to identify target phonemes. Analysing data from six subjects, we discern frequency bands that encapsulate differential information regarding production of vowels and consonants broadly, and more specifically nasals and semivowels. Subsequent spatial localization analysis reveals the underlying cortical regions responsible for different phoneme categories. Anatomical locations along with their respective frequency bands act as prospective feature sets for machine learning classifiers. We demonstrate this classification ability in a preliminary language reconstruction task and show an average word classification accuracy of 30.6% ($p<0.001$).

Index Terms: speech production, brain-computer interface, neural signal frequency bands

1. Introduction

Neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS) restrict an individual's potential to fully engage with their surroundings by hindering their communication abilities. Brain-Computer Interfaces (BCI) have long been envisioned to assist such patients as they bypass the affected pathways and directly translate neural recordings into text or speech output. These devices are trained to generate appropriate models of a subject's brain, and then classify and translate neural signals into commands.

The process of said translation consists of several steps. Initially, data is preprocessed to remove artifacts and noise, and to isolate underlying features in the signal relevant to speech production. Machine learning classifiers are then trained to identify target letters or phonemes at each time point. These classifications are then entered into a temporal process model which incorporates natural language information and smooths prediction output into final text [1].

However, practical implementation of this technology has been hindered by limitations in the speed and accuracy of existing systems [2]. The most widely used system utilizes P300 evoked response potentials (ERP) to spell out text by identifying individual attended characters. However, despite significant work in optimizing this system [3, 4], inherent limitations in its design result in slow typing speeds and a cumbersome training process [5]. More recently, systems using motor imagery to control a cursor or choose between tasks have been proposed [6]. However, these systems suffer from similar constraints as the "P300 spellers" and are significantly slower than spoken communication.

To address these shortcomings, several recent studies have proposed using electrocorticography (ECoG) and local field potential (LFP) signals for classification and reconstruction of individual phonemes and acoustic features [7, 8]. These invasive approaches provide superior signal quality as neural recordings are taken from directly on top of the cortex or within the cortical layer, thus capturing single cell events with high temporal and spatial accuracy. However, only two studies have attempted translation to continuous phoneme sequences using invasive neural data [9, 10]. Despite their reported higher translation speeds, their evaluations are limited to reduced dictionaries (10-100 words). Other design choices meant to enhance phoneme classification capitalize on prior knowledge of the output strings, hindering their generalization to unmodified, naturalistic speech.

An important challenge to decoding naturalistic speech lies in identification of the appropriate inputs for a classificatory scheme. Part of this endeavor is parsing features from neural signals relevant to the production of various phonemes. Previous studies have used signal power in frequency bands up until high-gamma (70-150 Hz) to map unto underlying speech [9, 10]. We instead analyze a wider range of frequencies and illustrate that often differential information about consonants and vowels is encoded in bands higher than high-gamma. We additionally deduce cortical regions responsible for production of vowels, nasals, semivowels, and consonants in general.

2. Methods

2.1. Participants and speech stimuli

Subjects in this study were neurosurgical patients with implanted intracranial depth electrodes to discern seizure foci for potential treatment of epilepsy [11]. Implantation of electrodes in the relevant areas of the temporal, frontal, and parietal lobes was based on clinical need. Six subjects performed between 55 and 603 trials, repeating either words ("yes", "no") or phoneme

http://dx.doi.org/10.21437/Interspeech.2019-3092

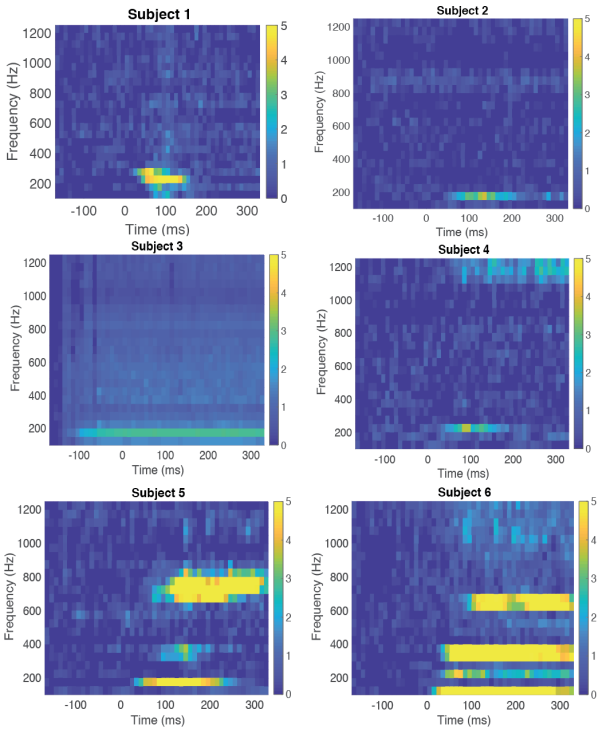

Figure 1: *Spectrograms for signal power change after consonant onset.*

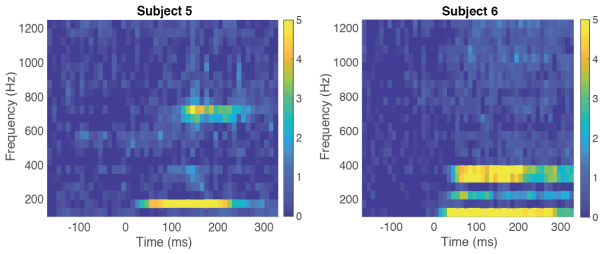

Figure 2: *Consonants by subjects 5 and 6 reproduced after removing signals related to following vowels.*

strings (singular vowels with or without preceding consonants). The number of phonemes per subject consequently varied from 8 (3 consonants, 5 vowels) to 16 (11 consonants, 5 vowels). The sampling rate of these recordings was 30 kHz. Before further processing, electrodes determined visually to have low signal-to-noise ratio (SNR) were removed.

2.2. Neural signal analysis

Each recording was divided into epochs (time windows) from -166.67 to 100 ms relative to onset of the speech. Labels were assigned respectively to the corresponding audio signal: [silence, speech]. In this paper, speech varyingly refers to consonants, vowels, nasals and semivowels.

The power per band is pre-processed by z-scoring and then downsampled to 100 Hz. These signals served as inputs to a linear classifier which was trained using early-stopping and coordinate descent methods. To additionally ensure that the classifier can correctly identify the silence after completion of the

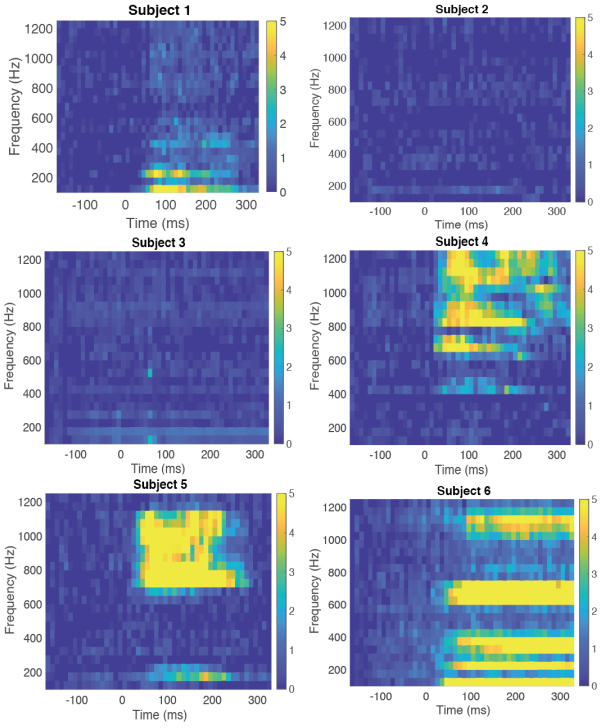

Figure 3: *Spectrograms for signal power change after vowel onset.*

phoneme string, we performed training over 100 ms post speech onset, but test the features captured by the classifier weights over 333.33 ms, since most trials end within this time period. The results illustrated in this paper, are obtained by z-scoring output of the classifier for each epoch using the mean and standard deviation of the -166.67 to -83.33 ms time interval.

All processing was done using the signal processing and STRFLab [13] toolboxes in MATLAB [14].

2.3. Signal localization

Localization of signals related to speech production was performed by aggregating standardized regression weights for microelectrode channels within brain regions for each subject [15]. Since electrode placement was based on clinical need, localization varied across subjects. For this analysis, only brain regions that had coverage in at least three subjects were noted. This constraint resulted in six regions: amygdala (subjects 2, 3, 4, and 6); anterior hippocampus (subjects 1, 2, and 6); entorhinal cortex (subjects 2, 3, 4, and 5); middle hippocampus (subjects 2, 3, and 4); medial orbitofrontal cortex (subjects 2, 3, 4, 5, and 6); and parahippocampal gyrus (subjects 1, 2, 3, and 4).

2.4. Preliminary speech reconstruction system

To demonstrate the ability to use these features in a BCI communication system, we constructed a preliminary system that uses these features to reconstruct spoken language from associated neural signals. Feature sets were determined by appending the frequency features defined here to averaged time domain signals in 10 ms time bins. A binary linear classifier was trained for each phoneme using leave-one-out cross-validation. A Gaussian mixture model was used to create probability distributions across the set of phonemes and these probabilities were

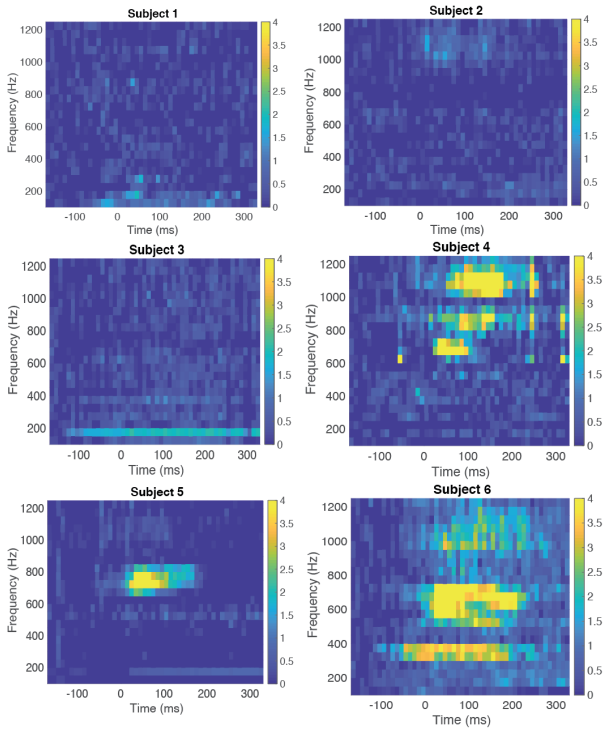

Figure 4: *Spectrograms for signal power change on incidence of vowels post consonants in a phoneme string / word.*

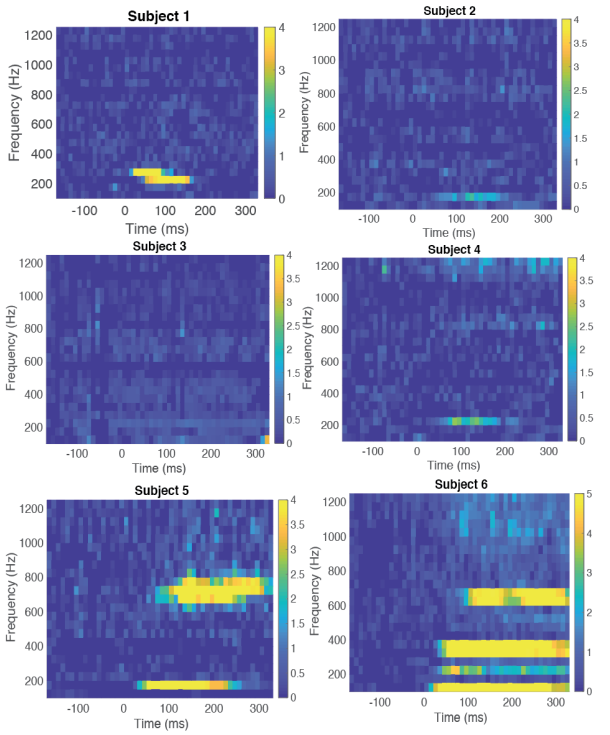

Figure 5: *Spectrograms for nasal onsets ("m, n").*

combined with a language model using a particle filtering algorithm developed for the P300 speller [12]. Performance was evaluated by comparing classification performance against the performance of the system using a randomized signal.

3. Results

In Fig. 1, we illustrate time frequency plots encapsulating information captured by different frequency bands during consonant production at the start of a word/phoneme string. As can be seen, the bands with high z-scores lie between 150 and 400 Hz for five of the six subjects. Due to low SNR, no visible patterns related to speech production emerged for subject 3.

Further, two subjects exhibit additional signals between 600 and 800 Hz. However, there is a time delay for the onset of these signals, indicating that they might be related to the vowels succeeding the consonant. This occurs due to variable lengths of phonemes across multiple utterances. We test for this by limiting our time window to primarily include only consonant production. Resultant Fig. 2 supports our hypothesis.

Similar analysis for vowels in Fig. 3 shows that for subjects 4,5, and 6, differential signals are predominantly contained in high frequency bands. However for subjects 1, 6, and, to a lesser extent, 5 there is also information contained in lower frequency bands. We allude to these findings later in our study of the underlying cortical regions responsible for vowel production.

Additionally, we identified bands responsible for encoding information about vowels when they succeed a consonant. Fig. 4 shows the emergence of primarily two regions in the high frequency bands across most participants. Subjects 4,5 and 6 exhibit high z-scores between 500 and 800 Hz, whereas 2 and 4 show similar structure between 1000 and 1200 Hz. Contrasted with Fig. 3, where we instead train on phoneme strings that begin with vowels, the frequency bands in Fig. 4 are narrower. Lastly, subjects 1 and 6 depict some low frequency information as before.

We sought to further identify bands that could aid in delineating amongst different categories of consonants. Since all speech by subjects 1,4,5, and 6 started with either a nasal or a semivowel phoneme, we study those in detail. Figs. 5 and 6 depict the localized nature of information about nasals and semivowels, respectively. We note that for subject 5 there again appears to be a delayed high frequency component in the nasal plot, which is likely a result of the following vowel as in Fig. 1.

After aggregating the standardized regression weights across electrodes within brain regions, we found that the highest average regression weights for vowel production were related to the medial orbitofrontal cortex (29% of total weight) and entorhinal cortex (24% of total weight). Weights were higher for the medial orbitfrontal cortex in high frequencies, while the amygdala had higher associated weights for low frequency signals. In consonant production, the amygdala (22% of total weight) and parahippocampal gyrus (22% of total weight) had the highest associated regression weights. Semivowels had higher weights associated with the parahippocampal gyrus and medial orbitofrontal cortex, while nasal consonants were more associated with signals in the entorhinal cortex.

In the speech reconstruction experiment, the average word accuracy across subjects was 30.6% with individual accuracies as high as 76% (Fig. 7). In each case, performance was significantly higher than that achieved after random permutation ($p<0.001$). Subject 2 had the lowest classification performance, likely because of the lack of defined power change for vowels. Interestingly, subject 3 had good classification performance despite not having well-defined power changes. This is possibly because time domain signals were effective in classification,

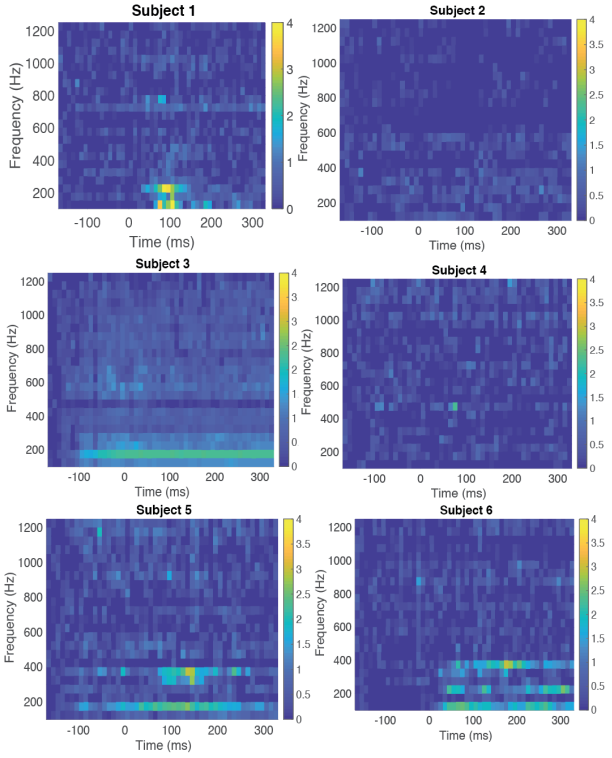

Figure 6: *Spectrograms for semivowel onsets ("y, l, r").*

which did not show up in the time-frequency analysis.

4. Discussion

Our time frequency plots show that information about consonants and vowels is often encoded in frequency bands higher than high-gamma. Consonants are primarily restricted to the 150-400 Hz bandwidth, while vowels are also encoded in higher frequencies, thus presenting differential input features for phoneme classification. More specifically, the illustrations in Fig. 4 provide further information aiding recognition of vowels that succeed consonants. Comparing it to Figs.1 and 2 one notices that for four out of the six subjects, the bands encoding either phoneme are fairly distinct. Additionally, while it appears that nasals and semivowels are both captured by similar frequency bands, our cortical analysis differentiates the regions associated with each. Lastly, similar subjects exhibit high z-scores for semivowels and low frequency band information for vowels. This we speculate could be due to their phonetic similarities and a potential differentiating feature for vowels would be the high frequency bandwidth that is visible in Fig. 3 for subjects 4, 5, and 6.

Signal localization showed that signals associated with different articulatory features occurred in different brain regions. The medial orbitofrontal cortex and entorhinal cortex were the areas most associated with vowel production, while the amygdala and parahippocampal gyrus were most associated with consonant production. Semivowels had higher association with channels located in the medial orbitofrontal cortex than nasal consonants. This similarity with vowels could be due to the fact that they are phonetically similar to vowel sounds. While the medial orbitofrontal cortex had associations in both low and high frequencies, it had higher associations when using higher

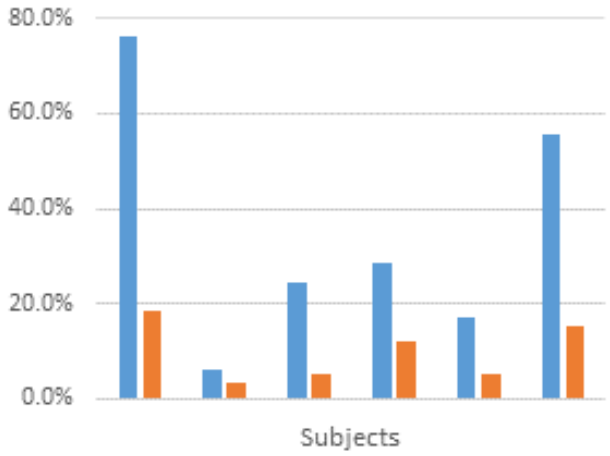

Figure 7: *Preliminary results for word accuracy during speech reconstruction for the six subjects (blue) compared to random signals (orange).*

frequency bands, particularly in subjects 4 and 5. This finding is consistent with previous studies performed using spike sorting that showed the medial orbitofrontal cortex contained sharply tuned neurons with high firing rates after speech onset [11]. It has also been observed that the superior temporal gyrus demonstrates broadly tuned units with lowering firing rates. While only two subjects in this study had superior temporal gyrus electrodes (subjects 1 and 2), signals from these channels were associated with lower frequency signals during vowel production in both subjects.

The focus of this study was to identify features that can be used for reconstructing language from neural signals. Frequency bands and brain regions identified in this analysis show that signals exist in multiple areas that are associated with different speech components. In the classification task, we showed that these features can be used to reconstruct spoken words. However, this analysis is retrospective using cross-validation, so it is unclear whether it will translate directly into a prospective system. Future work includes implementing these features to decode speech prospectively. Additional challenges also arise in a practical setting because overt speech can evoke auditory feedback signals that would not be available to ALS patients using a BCI communication device [16]. Hence studies should be performed to replicate current findings in a covert setting.

5. Conclusions

Different classes of phonemes are associated with changes in power spectra that vary across brain regions. This differential expression could potentially be used to reconstruct the produced speech from neural signals, which could be leveraged to create a real-time BCI communication device. Future work will consist of replicating these findings in a covert speech task and leveraging the features defined here in a classification task.

6. Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

7. References

[1] L. Comstock, A. Tankus, M. Tran, N. Pouratian, I. Fried, and W. Speier, "Developing a real-time translator from neural signals to text: An articulatory phonetics approach.," *Proc. Soc. Comput. Linguist*, vol. 2, no. 1, pp. 322–325, 2019.

[2] L. A. Farwell and E. Donchin "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials.," *Electroencephalogr. Clin. Neurophysiol*, vol. 70, no. 6, pp. 510-523, 1988.

[3] W. Speier, C. Arnold, N. Chandravadia, D. Roberts, S. Pendekanti, and N. Pouratian "Improving P300 spelling rate using language models and predictive spelling. ," *Brain- Computer Interfaces*, pp. 1-10, 2017.

[4] G. Townsend and V. Platsko "Pushing the P300- based braincomputer interface beyond 100 bpm: extending performance guided constraints into the temporal domain. ," *J. Neural Eng.*, vol. 13, no. 2, pp. 26024, 2016.

[5] J. E. Huggins, P. A. Wren and K. L. Gruis "What would brain-computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis. ," *Amyotroph. Lateral Scler.*, vol. 12, no. 5, pp. 318-324, 2011.

[6] D. J. McFarland, L. A. Miner T. M. Vaughan, and J. R. Wolpaw "Mu and beta rhythm topographies during motor imagery and actual movements. ," *Brain Topogr.*, vol. 12, no.3, pp. 177-186, 2000.

[7] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, and S. A. Shamma "Reconstructing Speech from Human Auditory Cortex ," *PLOS Biology*, vol. 10, no.1, pp. 1-13, 2012.

[8] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific Reports*, vol. 10, pp. 874, 2019.

[9] C. Herff, D. Heger, A. Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz "Brain-to-text: decoding spoken phrases from phone representations in the brain.," *Front. Neurosci.*, vol. 9, pp. 217, 2015.

[10] D. A. Moses, N. Mesgarani, M. K. Leonard, and E. F. Chang "Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity.," *J. Neural Eng.*, vol. 13, no. 5, pp. 56004, 2016.

[11] A. Tankus, I. Fried, and S. Shoham "Structured neuronal encoding and decoding of human speech features.," *Nat. Commun.*, vol. 3, pp. 1015, 2012.

[12] W. Speier, C. Arnold, A. Deshpande, J. Knall, and N. Pouratian "Incorporating advanced language models into the P300 speller using particle filtering.," *J. Neural Eng.*, vol. 12, pp. 046018, 2015.

[13] *STRFLab Toolbox,*, v1.45, Auditory Science Lab, Berkeley.

[14] *MATLAB and Statistics Toolbox,*, Release 2018a, Natick, MA: The Mathworks Inc., 2016.

[15] L. L. Nathans, F. L. Oswald, and K. Nimon "Interpreting Multiple Linear Regression: A Guidebook of Variable Importance.," *Practical Assessment, Research and Evaluation*, vol. 17, pp. 1-19, 2012.

[16] S. Martin, I. Iturrate, J. Millan, R. Knight, and B. N. Pasley "Decoding Inner Speech Using Electrocorticography: Progress and Challenges Toward a Speech Prosthesis.," *Front. Neurosci.*, vol. 12, pp. 422, 2018.