



# Evidence of Onset and Sustained Neural Responses to Isolated Phonemes from Intracranial Recordings in a Voice-based Cursor Control Task

Kevin Meng<sup>1</sup>, Seo-Hyun Lee<sup>2</sup>, Farhad Goodarzy<sup>3</sup>, Simon Vogrin<sup>3</sup>, Mark J. Cook<sup>1,3,5</sup>,  
Seong-Whan Lee<sup>2,4</sup>, David B. Grayden<sup>1,3,5</sup>

<sup>1</sup> Department of Biomedical Engineering, The University of Melbourne

<sup>2</sup> Department of Brain and Cognitive Engineering, Korea University

<sup>3</sup> Department of Medicine, St Vincent's Hospital, The University of Melbourne

<sup>4</sup> Department of Artificial Intelligence, Korea University

<sup>5</sup> Graeme Clark Institute for Biomedical Engineering, The University of Melbourne

ksmeng@student.unimelb.edu.au, {seohyunlee, sw.lee}@korea.ac.kr,  
{goodarzy, vogrin, markcook, grayden}@unimelb.edu.au

## Abstract

We developed a voice-based, self-paced cursor control task to collect corresponding intracranial neural data during isolated utterances of phonemes, namely vowel, nasal and fricative sounds. Two patients implanted with intracranial depth electrodes for clinical epilepsy monitoring performed closed-loop voice-based cursor control from real-time processing of microphone input. In post-hoc data analyses, we searched for neural features that correlated with the occurrence of non-specific speech sounds or specific phonemes. In line with previous studies, we observed onset and sustained responses to speech sounds at multiple recording sites within the superior temporal gyrus. Based on differential patterns of activation in narrow frequency bands up to 200 Hz, we tracked voice activity with 91% accuracy (chance level: 50%) and classified individual utterances into one of five phonemes with 68% accuracy (chance level: 20%). We propose that our framework could be extended to additional phonemes to better characterize neurophysiological mechanisms underlying the production and perception of speech sounds in the absence of language context. In general, our findings provide supplementary evidence and information toward the development of speech brain-computer interfaces using intracranial electrodes.

**Index Terms:** phoneme recognition, intracranial electrodes, speech onset, sustained speech, brain-computer interfaces

## 1. Introduction

Intracranial recording modalities, such as electrocorticography (ECoG) and stereotactic electroencephalography (SEEG), may be used to restore direct communication at a reasonable speed for patients who lost their ability to speak [1]. Recent brain-computer interface (BCI) systems aim to convert silent speech processes from intracranial recordings into useful commands in the environment [2, 3]. In these systems, model training relies on supervised learning using either attempted speech [2] or overt speech [3]. Labels may be phonemes in isolation [4, 5], phonemes in context [6, 7, 8], syllables or words [2, 9, 10, 11], audio features [3, 12, 13] or articulatory parameters [14]. Furthermore, language models may be exploited to improve decoding results [15].

Understanding neurophysiological mechanisms underlying the production and perception of speech is essential to improve speech decoding performance. Interestingly, the spatiotemporal resolution of intracranial electrodes was shown to be sufficient to capture neural activity that reflects language processing at a phonetic level. For example, using ECoG grids, Chartier et al. [16] proposed that the ventral sensorimotor cortex is organized in clusters of neurons that are sensitive to place of articulation (e.g., coronal, labial, dorsal) from an overt reading task, while Hamilton et al. [17] suggested that the superior temporal gyrus is parcellated into neuronal populations that are sensitive to manner of articulation (e.g., vowel, nasal, fricative, plosive) during perception of natural speech. Based on the high-gamma frequency component of the intracranial recordings, which is known to correlate with multi-unit firing rates [18], the latter study also found that neuronal clusters may be non-selective and may encode either speech onset or sustained speech.

In this study, we collected intracranial signals using SEEG depth electrode arrays, which penetrate into deeper structures of the brain beyond cortical surfaces. Study participants produced isolated phonemes in the absence of any language context and perceived their own voice in an engaging, voice-based, self-paced cursor control task. Following voiced-based cursor control, we conducted post-hoc analyses of the neural data. Despite the lack of control on electrode placement, we hypothesized the existence of discriminative features in our dataset to detect the occurrence of non-specific speech sounds and classify individual utterances into one of five phonemes. Thus, we report patient-specific outcomes in terms of neural feature contributions and prediction accuracy. In conclusion, we interpret our findings from a neurophysiological perspective and discuss potential implications for speech BCI systems.

## 2. Materials and Methods

### 2.1. Study participants

Two male patients, native speakers of Australian English with intractable epilepsy, were implanted with 11 and 15 SEEG depth electrode arrays (DIXI Medical, France), consisting of 10-18 contacts per array, for a total of 132 and 226 contacts. The placement of the arrays, mostly in the right temporal lobe for both patients, was based on the requirements of clinical

evaluation. Intracranial signals were sampled at 5 kHz using a Neuvo amplifier and Profusion EEG software (Compumedics Ltd, Australia). Audio signals were sampled at 16 kHz using a unidirectional Snowball iCE microphone (Blue Microphones, USA). Both data streams were synchronized with event markers using StimTracker (Cedrus Corporation, USA). The study was approved by the Human Research Ethics Committee at St Vincent’s Hospital Melbourne (SVHM HREC-A 267/20).

## 2.2. Voice-based cursor control

Inspired by the Vocal Joystick application [19], we processed the microphone input in real time to estimate the formant frequencies of sustained vowel sounds for continuous cursor movement in the two-dimensional plane. Motivated by an intracortical BCI system for point-and-click control [20], we created an additional degree of freedom by detecting fricative sounds for click selections (see **Figure 1A**). Predictions were made every 16 ms using custom Python code (version 3.9 with the following packages: `scipy` [21], `parselmouth` [22] and `scikit-learn` [23]). First, the ratio of energy contained in the high-frequency band (4000-8000 Hz) over the low-frequency band (0-4000 Hz) was estimated. If this ratio exceeded an arbitrary threshold (for fricative sounds), a click selection was produced. Second, the energy contained in the expected voicing range (75-300 Hz) was estimated. If this value remained below another arbitrary threshold (for vowel and nasal sounds), nothing happened. If it exceeded the threshold, instantaneous formant frequencies were estimated via linear predictive coding [24]. The first two formants were used as model inputs into a pre-trained Gaussian Mixture Model (GMM) decoder to predict one of four possible directions. All audio features were estimated using a window size of 50 ms.

## 2.3. Model calibration and data collection

The cursor could be moved in four directions within the two-dimensional plane based on which vowel was uttered. To train the GMM decoder with their own voice, study participants produced one sustained utterance of a low vowel (/a/), a high-front vowel (/i/), a mid-back vowel (/o/), and a nasal sound (/m/) (see **Figure 1B**). Each sound was assigned to one of the four directions for cursor movement. Following this quick calibration procedure, the framework for the main experiment

consisted of a visual display with a cursor and a target (see **Figure 1C**). For at least 120 seconds, study participants were instructed to move the cursor toward the target, which could be selected by means of an affricate (/tʃ/) sound.

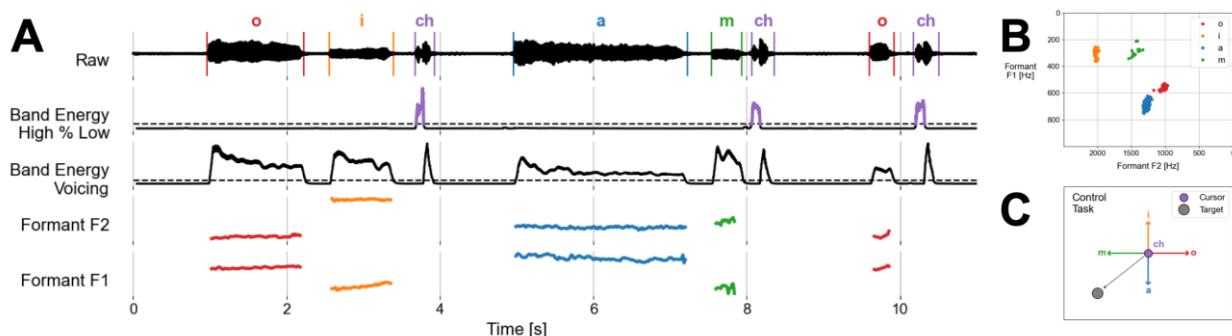
## 2.4. Identification of discriminative neural features

To identify neural features that contained information about non-specific speech sounds, Pearson correlation coefficients were repeatedly computed to quantify the relationship between the transformed audio and neural data. On one hand, binary time series (1: non-specific speech, 0: silence) were created from the raw audio signal. For “onset” analyses, segments of “1” values were limited to 100 ms. For “sustained” analyses, segments of “1” values were not limited. On the other hand, continuous-valued time series (values: band energy over time, window size of 100 ms) were created from the raw neural signal at each recording site. A bipolar referencing scheme was applied to emphasize local neural activity [25] and to minimize potential artifacts including acoustic contamination [26]. Narrow frequency bands were selected from 0-200 Hz in steps of 4 Hz. To identify neural features that contained information about specific phonemes, the methods described above were applied by focusing on each phoneme separately and ignoring the other phonemes. In other words, the creation of binary time series from the raw audio signal differed in that they contained “1” values for the phoneme of interest and “0” values for all other phonemes and silences. The creation of continuous-valued time series from the raw neural signal remained identical.

## 2.5. Voice activity tracking and phoneme classification

To track voice activity from neural signals, we ranked neural features by their correlation coefficient from the “sustained” analyses. Binary support vector machine (SVM) decoders were trained using the best neural features as model inputs. Data points were extracted every 16 ms. For this classification task, labels were either “non-specific speech” or “silence”.

To classify individual utterances into one of five phonemes from neural signals, we ranked neural features by their correlation coefficient, but this time from both the “onset” and “sustained” analyses. Five-class SVM decoders were trained using the best neural features as inputs. In this case, labels were “/a/”, “/i/”, “/o/”, “/m/” or “/tʃ/”.



**Figure 1. Overview of the voice-based cursor control task.** (A) The microphone input was processed in real time to control a point-and-click application using five pre-selected phonemes: /a/, /i/, /o/, /m/ and /tʃ/. Audio features, such as the ratio of high-band over low-band energy, the band energy in the voicing range and the first two formant frequencies, were extracted every 16 ms using a window size of 50 ms. (B) Examples of the four different vowel and nasal sounds form clusters in the formant space. These phonemes could be accurately classified using a Gaussian Model Mixture (GMM) decoder. (C) Study participants were instructed to move the cursor toward a target using vowel and nasal sounds, and produce a click selection by means of an affricate sound.

After splitting our dataset once in chronological order using a 75:25 train-test ratio to reflect hypothetical real-time brain control scenarios, we progressively increased the number of inputs and reported the corresponding accuracies averaged over 10 repetitions for both classification tasks, namely voice activity tracking and phoneme classification.

### 3. Results

#### 3.1. Neural features for non-specific speech sounds

To find neural features for non-specific speech sounds, binary time series created from the raw audio signal contain “1” values for all phonemes (see **Figure 2A-B**). Heatmaps of correlation values at different frequency bands (y-axis) for all individual contacts (x-axis) along each depth array (subplot) are shown for the first study participant only (see **Figure 2C-E**).

Two depth electrode arrays in the right superior temporal gyrus (STG) – H in the middle STG, S in the posterior STG – showed overall higher correlation values compared to the other arrays in other brain areas. Interestingly, high individual values were spread across various low and high frequency bands for the “onset” analyses but could only be observed in localized high frequency bands for “sustained” analyses. For example, focusing on array S, speech onset seems to be encoded in the vicinity of contacts S3-S5 at around 4-12 Hz and 60-120 Hz. In contrast, sustained speech seems to be encoded near contacts S4-S8 (spatially broader) at around 100-110 Hz (spectrally more localized).

#### 3.2. Neural features for individual phonemes

To find neural features for individual phonemes, binary time series created from the raw audio signal contain “1” values only for the phoneme of interest in the desired analysis (see **Figure 3A-B**). Heatmaps of correlation values at different frequency

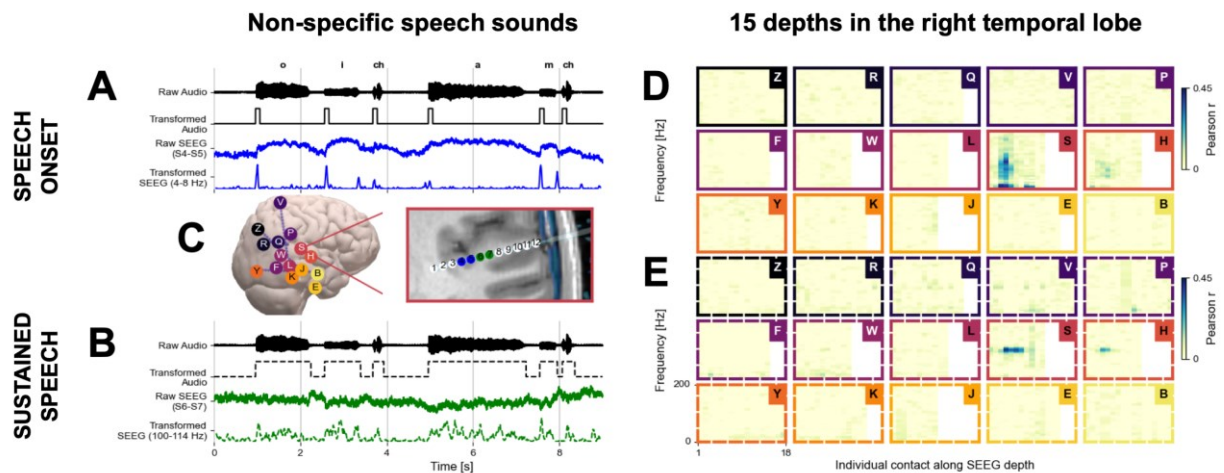
bands (y-axis) for all individual contacts along selected depth arrays (x-axis) for different phonemes separately (subplot) are shown for the first study participant only (see **Figure 3C-F**).

Compared to the heatmaps for non-specific sounds (both onset and sustained), differential patterns of correlation values emerged from phoneme-specific analyses, thereby emphasizing features in the spatial and spectral dimensions that may encode information about specific phonemes or group of phonemes. For example, in the “sustained” analyses for array H, three phonemes (/o/, /m/, /a/) showed high correlation values at 90-100 Hz between contacts H3-H7 (see **Figure 3F**). However, differences could be observed in the spatial dimension, which suggest phoneme-specific encoding around individual contacts: /o/ at H4-H5 and H7-H8, /m/ at H5-H6, and /a/ at H6-H7. In contrast, in the “sustained” analyses for array S, two phonemes (/o/, /m/) showed high correlation values at 90-110 Hz between contacts S4-S8 (see **Figure 3D**). Here, differences could be observed in the spectral dimension, which suggest phoneme-specific encoding by identical neuronal populations, but around different frequencies: /o/ at 90-100 Hz, /m/ at 100-110 Hz.

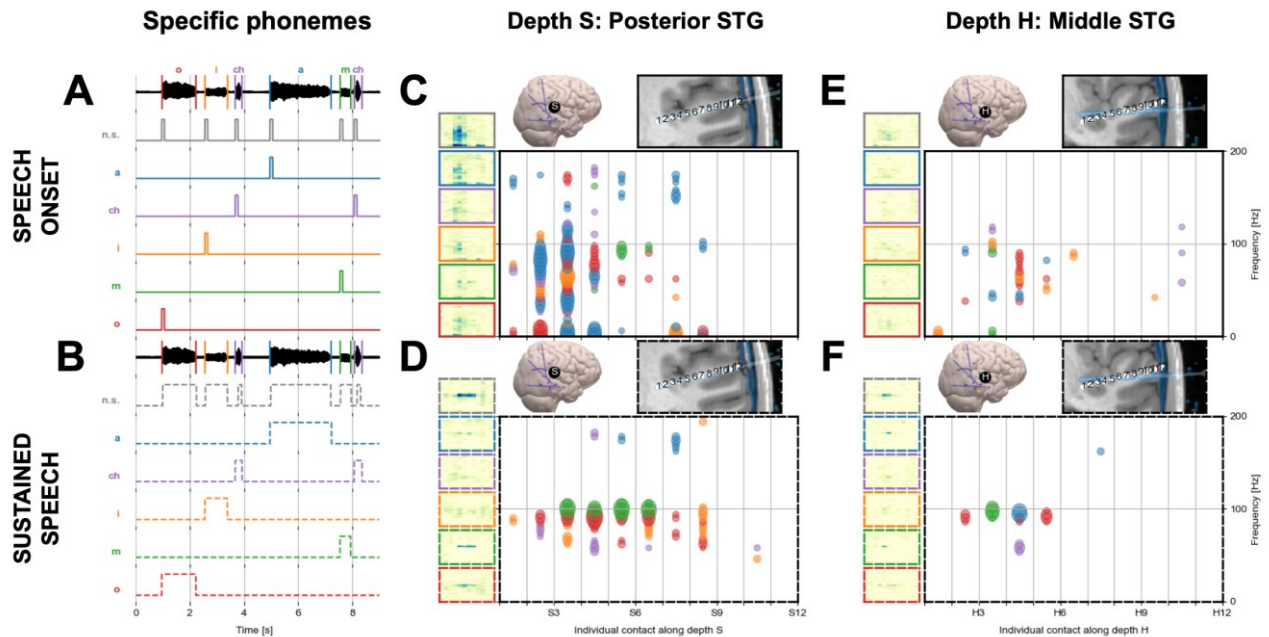
#### 3.3. Voice activity tracking and phoneme classification

Using the best neural features as inputs in SVM decoders, voice activity tracking was achieved with up to 91% accuracy (chance level: 50%) and five-class phoneme classification was achieved with up to 68% accuracy (chance level: 20%).

For both classification tasks, prediction accuracy increased with the number of input features, ranked by their correlation coefficient in “onset” and “sustained” analyses, as described in Section 2.5. For example, prediction accuracy for voice activity tracking quickly reached 85% using the best 10 neural features, then plateaued with a peak value of 91% at about 200 neural features. The curve for phoneme classification followed a similar trend with lower values.



**Figure 2. Neural features for the detection of non-specific speech sounds.** (A) The audio signal is transformed into a binary time series representing speech onsets (100 ms). The intracranial signal is pre-processed with bipolar reference and transformed into its power spectral density in a narrow frequency band. This example is for channels S4-S5 at 4-8 Hz (blue trace). Transformed audio and neural signals were used to compute Pearson correlation coefficients. (B) The binary time series represents sustained speech. The intracranial example is for channels S6-S7 at 100-114 Hz (green trace). (C) In total, 15 SEEG depth electrode arrays (colored circles) were implanted in the right hemisphere of one study participant. The pink-framed rectangle shows individual channels along array S. The location of channels S4-S5 (blue circles) and S6-S7 (green circles) is emphasized. (D-E) Heatmaps represent correlation values for each array (subplot) at different recording sites (x-axis) and narrow frequency bands (y-axis) for onset (solid frame) and sustained (dashed frame) analyses. In each subplot, dark colors indicate high values. Colored circles in C correspond to colored frames in D-E.



**Figure 3. Neural features for the classification of specific phonemes.** (A-B) Binary time series from the audio signal for onset (top half) and sustained (bottom half) correlation analyses. (C-D) Analyses for array S in the posterior superior temporal gyrus (STG). Left figures: Heatmaps of correlation values for each phoneme. Right figure: Summary of phonemes with maximum correlation value at individual spatial (x-axis) and spectral (y-axis) locations across the heatmap if the maximum value exceeds a threshold. Colors represent phonemes. Marker size represents correlation value. (E-F) Same analyses for array H in the middle STG.

#### 4. Discussion

The first aim of this study was to elicit isolated phonemes and simultaneously acquire intracranial recordings. Closed-loop voice-based cursor control was successfully achieved by study participants, which generated a unique audio-SEEG dataset of phonemes in isolation and silences of different durations. The second aim was to confirm the existence of neural features in our dataset to track non-specific speech and classify individual utterances into one of five possible phonemes.

In general, our results are consistent with previous studies of the STG using intracranial signals [17, 27]. Our analyses add further evidence about the parcellation of this brain region into neuronal populations that encode either speech onset or sustained speech, but also non-specific or specific sounds. In addition, we offer a fresh perspective into this theory by (1) recording from deeper structures using SEEG depth electrodes, as opposed to ECoG surface grids, (2) analyzing phonemes in isolation, not in context, and (3) investigating narrow frequency bands instead of broad high-gamma bands. More specifically, while we expected to observe spatial patterns to discriminate between phonemes (see **Figure 3F**), we did not expect to see the emergence of spectral patterns, which suggest that the same neuronal populations may encode information about different phonemes around close frequency values (see **Figure 3D**).

Using two- and five-class linear classifiers with a limited number of features, we obtained prediction accuracies that were well above chance level. This indicates that our dataset could be further exploited through deeper analyses, such as tracking formant frequencies of vowel sounds and conducting pseudo-prospective simulations to reconstruct cursor trajectories from intracranial signals exclusively.

Our study was limited by the small number of participants and the lack of control on electrode placement due to the intrinsic nature of intracranial research in clinical settings. It is likely that onset and sustained responses observed in our dataset reflect neural processes related to perception of own voice rather than intention to produce overt speech. In the future, replicating the experiment with better coverage of speech-related areas would help to disentangle the neurophysiological interactions between speech production and perception.

To conclude, our findings provide supplementary evidence for the development of practical BCI applications and the feasibility of speech BCI systems. On one hand, we could create additional degrees of freedom through the reliable detection of sustained sonorant or short fricative sounds. On the other hand, a better understanding of neural processes in speech-related brain areas may accelerate the calibration and optimize the performance of continuous speech decoding models.

#### 5. Acknowledgements

We thank the clinical team at St Vincent's Hospital Melbourne for their support, as well as the participants for dedicating their time for voluntary research during their admission.

KM, FG, SV, MJC, DBG were supported by the NHMRC Project Grant 1148005. SHL and SWL were supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2017-0-00451, Development of BCI based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning; No. 2021-0-02068, Artificial Intelligence Innovation Hub).

## 6. References

- [1] E. F. Chang and G. K. Anumanchipalli, "Toward a speech neuroprosthesis," *Jama*, vol. 323, no. 5, pp. 413-414, 2020.
- [2] D. A. Moses et al., "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria," *New England Journal of Medicine*, vol. 385, no. 3, pp. 217-227, 2021.
- [3] M. Angrick et al., "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity," *Communications Biology*, vol. 4, no. 1, pp. 1-10, 2021.
- [4] S. Ikeda et al., "Neural decoding of single vowels during covert articulation using electrocorticography," *Frontiers in Human Neuroscience*, vol. 8, p. 125, 2014.
- [5] E. C. Leuthardt et al., "Using the electrocorticographic speech network to control a brain-computer interface in humans," *Journal of Neural Engineering*, vol. 8, no. 3, p. 036004, 2011.
- [6] E. M. Mugler et al., "Direct classification of all American English phonemes using signals from functional speech motor cortex," *Journal of Neural Engineering*, vol. 11, no. 3, p. 035015, 2014.
- [7] C. Herff et al., "Brain-to-text: Decoding spoken phrases from phone representations in the brain," *Frontiers in Neuroscience*, vol. 9, p. 217, 2015.
- [8] 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," *Journal of Neural Engineering*, vol. 13, no. 5, p. 056004, 2016.
- [9] S. Martin et al., "Word pair classification during imagined speech using direct brain recordings," *Scientific Reports*, vol. 6, p. 25803, 2016.
- [10] G. Milsap, M. Collard, C. Coogan, Q. Rabbani, Y. Wang, and N. E. Crone, "Keyword spotting using human electrocorticographic recordings," *Frontiers in Neuroscience*, vol. 13, p. 60, 2019.
- [11] T. Proix et al., "Imagined speech can be decoded from low- and cross-frequency intracranial EEG features," *Nature Communications*, vol. 13, no. 1, pp. 1-14, 2022.
- [12] S. Martin et al., "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in Neuroengineering*, vol. 7, p. 14, 2014.
- [13] M. Angrick, C. Herff, G. D. Johnson, J. J. Shih, D. J. Krusienski, and T. Schultz, "Speech spectrogram estimation from intracranial brain activity using a quantization approach," in *Proceedings of INTERSPEECH*, 2020, pp. 2777-2781.
- [14] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493-498, 2019.
- [15] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nature Communications*, vol. 10, no. 1, pp. 1-14, 2019.
- [16] J. Chartier, G. K. Anumanchipalli, K. Johnson, and E. F. Chang, "Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex," *Neuron*, vol. 98, no. 5, pp. 1042-1054. e4, 2018.
- [17] L. S. Hamilton, E. Edwards, and E. F. Chang, "A spatial map of onset and sustained responses to speech in the human superior temporal gyrus," *Current Biology*, vol. 28, no. 12, pp. 1860-1871. e4, 2018.
- [18] N. E. Crone, A. Sinai, and A. Korzeniewska, "High-frequency gamma oscillations and human brain mapping with electrocorticography," *Progress in Brain Research*, vol. 159, pp. 275-295, 2006.
- [19] S. Harada, J. A. Landay, J. Malkin, X. Li, and J. A. Biles, "The vocal joystick: Evaluation of voice-based cursor control techniques," in *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, 2006, pp. 197-204.
- [20] C. Pandarinath et al., "High performance communication by people with paralysis using an intracortical brain-computer interface," *Elife*, vol. 6, p. e18554, 2017.
- [21] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261-272, 2020.
- [22] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1-15, 2018.
- [23] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
- [24] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 129-134, 1993.
- [25] M. R. Mercier et al., "Evaluation of cortical local field potential diffusion in stereotactic electro-encephalography recordings: a glimpse on white matter signal," *Neuroimage*, vol. 147, pp. 219-232, 2017.
- [26] P. Roussel et al., "Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception," *Journal of Neural Engineering*, vol. 17, no. 5, p. 056028, 2020.
- [27] A. Tankus, I. Fried, and S. Shoham, "Structured neuronal encoding and decoding of human speech features," *Nature Communications*, vol. 3, no. 1, pp. 1-5, 2012.