2023 11th International IEEE/EMBS Conference on Neural Engineering (NER) | 978-1-6654-6292-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/NER52421.2023.10123716

Discrimination of Overt, Mouthed, and Imagined Speech Activity using Stereotactic EEG

P. Z. Soroush[1], S. Y. Cole[1], C. Herff[2], S. K. Ries[3], J. J. Shih[4], T. Schultz[5], and D. J. Krusienski[1]

***Abstract*— Recent studies have demonstrated that it is possible to decode and synthesize acoustic speech directly from intracranial measurements of brain activity. A current major challenge is to extend the efficacy of this decoding to imagined speech processes toward the development of a practical speech neuroprosthesis for the disabled. The present study used intracranial brain recordings from participants that performed a speaking task consisting of overt, mouthed, and imagined speech trials. In order to better elucidate the unique neural features that contribute to the discrepancies between overt and imagined model performance, rather than directly comparing the performance of speech decoding models trained on respective speaking modes, this study developed and trained models that use neural data to discriminate between pairs of speaking modes. The results further support that, while there exists a common neural substrate across speech modes, there are also unique neural processes that differentiate speech modes.**

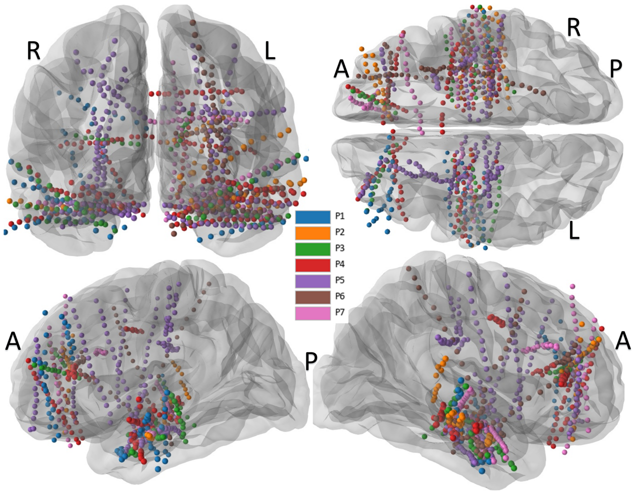


Fig. 1: The combined sEEG depth electrode (channel) locations of the 7 participants from different perspectives using an averaged brain model.

I. Introduction

Speech is the first and foremost modality of human interpersonal communication. Brain-Computer Interfaces (BCIs) that decode and synthesize speech could dramatically improve life for individuals unable to speak due to injury or disease. Invasive measurements of brain activity using electrocorticography (ECoG) [1] or stereotactic electroencephalography (sEEG) [2] have recently shown promise for developing such speech BCIs [3], [4], [5], [6], [7].

For those who have lost the ability to speak, the objective is to translate neural processes during *imagined* speech to acoustic speech. However, the lack of behavioral output during imagined speech makes it extremely challenging to design an effective decoding model [8], [9]. To overcome this challenge, studies often employ neural processes or behavioral output from *overt* or *mouthed* (i.e., inaudible articulations without vocalization) speech as a surrogate to study associated neural activity [3], [10] or to train decoding models [5], [7], [11] for imagined speech applications.

While these studies have shown substantial promise, there are clear limitations to using overt speech surrogates for training imagined-speech decoding models. This may be due, in part, to the unique brain regions activated during overt, mouthed, and imagined speech, and the differences between the neural features extracted from these regions [8], [12], [13], [14]. In order better elucidate the unique neural features that contribute to the discrepancies between overt and imagined model performance, rather than directly comparing the performance of speech decoding models trained on respective speaking modes, the present study developed and trained models that used neural data to discriminate between pairs of speaking modes.

II. Methodology

A. Participants and Electrode Locations

sEEG data were collected from 7 native English-speaking participants being monitored as part of treatment for intractable epilepsy at UCSD Health. The demographic information of the participants is provided in Table I. The study design was approved by the Institutional Review Boards of Virginia Commonwealth University and UCSD Health, and informed consent was obtained for experimentation with human subjects. The locations of sEEG electrodes were determined solely based on the participants' clinical needs. A subset of the implanted electrodes for each participant was determined to be in or adjacent to brain regions associated with speech and language processing. Fig. 1 shows the depth electrode locations for the 7 participants, with sEEG electrode (channel) counts provided in Table I.

B. Experimental Design and Data Collection

The experimental setup and trial sequence structure are depicted in Fig. 2. For each trial sequence, a sentence

*This work was supported by NSF 2011595/1608140 and BMBF 01GQ1602.

[1] P. Z. Soroush, S. Y. Cole, and D. J. Krusienski are with Virginia Commonwealth University, Richmond, VA, USA zanganehp@vcu.edu

[2] C. Herff is with the University of Maastricht, Maastricht, Netherlands

[3] S. Ries is with San Diego State University, San Diego, CA, USA

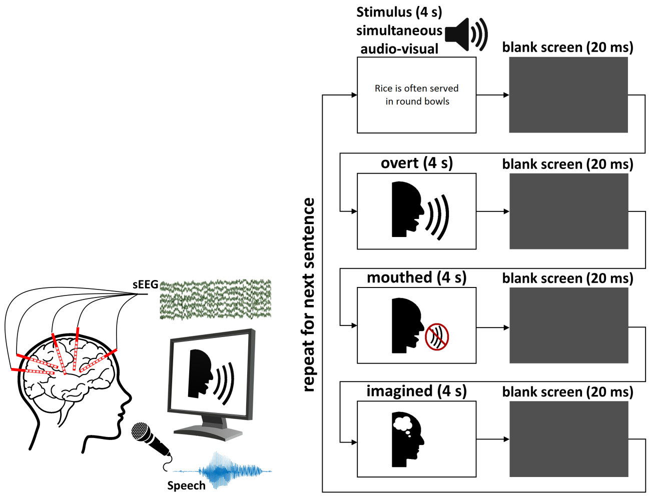
[4] J. J. Shih is with UCSD Health, San Diego, CA, USA

[5] T. Schultz is with the University of Bremen, Bremen, Germany

Participant	Gender	Age	# Recorded	# Excluded Channels	# Trials
P1	Male	25	90	5	25
P2	Male	60	70	3	50
P3	Male	32	80	5	50
P4	Female	42	175	11	50
P5	Male	21	232	17	50
P6	Male	22	94	5	50
P7	Male	31	108	8	50

TABLE I: Demographic information, numbers of sEEG channels, and unique trial sequences for each participant.

was displayed on a computer monitor and simultaneously narrated via computer speakers for a 4-second interval. While the acoustic speech and sEEG signals were simultaneously recorded, the participants were visually prompted by a sequence of icons as cues to (1) speak the sentence audibly (overt), (2) inaudibly articulate the sentence (mouth), and (3) imagine speaking the sentence without articulating or vocalizing (imagine). This structure was repeated for 50 unique sentences [14]. Each icon prompt and participant response during the task is referred to as a single *trial*. The stimuli were presented and synchronized with the sEEG recordings using Presentation® software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com).



(a) Experimental setup.

(b) Trial sequence structure.

Fig. 2: sEEG and acoustic signals were simultaneously recorded as the participant performed the task as prompted by icons presented on a monitor.

The sEEG electrodes were referenced to a pair of subdermal needle electrodes in the scalp and digitized at 1,024 Hz. The audio signal was recorded via an external microphone and digitized at 44,100 Hz. The audio recordings of the overt trials were used to label the data as *speech* or *non-speech* in 10-ms non-overlapping segments. The 10-ms frame length was chosen to capture the relevant temporal dynamics of speech activity for eventual closed-loop implementation. These labels were used to define surrogate labels for the mouthed and imagined trials [14].

C. Data Pre-processing and Feature Extraction

All sEEG data were visually inspected for noisy or anomalous channels, as well as was analyzed for any potential audio contamination or perceptual information [14], [15]. The number of channels excluded as a result of this screening are reported in Table I. The remaining raw sEEG channels were re-referenced using the Laplacian method [16].

The narrow-band power of each sEEG channel was computed in four conventional frequency bands: theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), and broadband gamma (70-170 Hz), based on prior studies [14], [17]. To extract the spectral features, using the labeled 10-ms frames from the audio signals, the sEEG channels over a specified temporal window around each audio frame were zero-phase filtered for using a sixth-order Butterworth filter for each frequency band. A 210-ms window length was chosen (corresponding to 200 ms before the frame to the end of the frame). An additional notch filter from 118-122 Hz was applied to broadband gamma to suppress the second harmonic of the 60 Hz line noise. The features were computed every 10 ms as the natural logarithm of the signal energy over 210 ms, representing 10 ms overlapping the audio frame and 200 ms prior to the frame to emulate a causal design. This causal design was selected to ensure decoding activity related to speech production rather than perception for the future implementation in real-time, closed-loop applications.

The features from each included channel were concatenated to form the feature vector (# channels × 21 features × # frequency bands - representing spatial, temporal, and spectral information, respectively) for the decoding models.

D. Mode-Discriminating Models

Mode-Discriminating (MD) models were developed to investigate the similarities and differences across speech modes beyond basic *speech/non-speech* gating. For each participant and mode, approximately one tenth of the trials were randomly selected and reserved for evaluating the models, herein referred to as *MD test trials*. The data from the remaining trials, referred to as *MD training trials*, were used to perform channel selection and model training as subsequently described.

1) Logistic Regression Model and Classification Evaluation: All models were designed using logistic regression with L1 regularization and were specific for each participant and comparison [17], [18]. Due to the difference between the amount of data for each class in some of the models, and for consistency, the performance of all models was evaluated using balanced accuracy (i.e., the average of the recalls of the classes). To establish the chance-level classification for each model, a randomization test was performed where all labels were randomly shuffled and the cross-validation process was repeated for 1,000 separate randomizations of the labels.

All significance tests were performed using a Benjamini-Hochberg corrected Wilcoxon signed-rank test. The resulting p-value (p) is reported for each respective test.

2) Channel-Selection: Channel selection for the MD models was performed by comparing the relative performance of single channels for speech activity detection (i.e., binary classification of speech versus non-speech segments) using logistic regression on the *MD training trials*. For each mode, the decoding performance was evaluated using a 10-fold, non-shuffled cross-validation. To prevent from training bias, during the cross-validation process, the training data (partitioned from the independent validation and test data) was normalized to zero mean and unit variance, and the same normalization parameters were applied to the validation (one tenth of the data in the nine training folds) and test data. The validation data was used to optimize the hyperparameters of the training models, while the test data was solely used to obtain the performance of the trained models for each fold.

For each participant, mode, and fold in the cross-validation process, the mean plus one standard deviation of the balanced accuracy of all channels was determined as the threshold for the fold, which were aggregated to form a distribution of thresholds over the ten folds of the cross-validation process. Additionally, for each channel, the distribution of balanced accuracies over the 10 folds of the cross-validation process was computed. The distribution of balanced accuracies were used to identify channels with discrepant performance across each pairwise combination of modes.

For each mode pair, the threshold computed for the first mode was used to select channels that performed significantly better than this threshold ($p < 0.05$) for the first mode and performed significantly below this threshold ($p < 0.05$) for the second mode. This way, channels uniquely relevant to only one mode in the pair were selected.

3) Mode-Discriminating Models: For each pairwise combination of speech modes, the *MD training trials* and *MD test trials* were respectively parsed into *speech* (combination of *speech* segments of the two modes) and *non-speech* (combination of *non-speech* segments of the two modes) conditions. The training data was then normalized to zero mean and unit variance, and the same normalization parameters were applied to the test data. Next, for each selected channel and each *speech* and *non-speech* condition, a logistic regression model was trained and tested on the training and test data, respectively. The MD models are labeled as (Mode A)-(Mode B). For each mode pair and condition, the average of the performance of the selected channels was used to compare the differences between the brain activity during the two modes.

III. Results

Overt, mouthed, and imagined speech modes can be compared based on their respective degree of behavioral output [14]. In the subsequent paired comparisons, the mode in the pair having the higher behavioral output and the mode with lower behavioral output will be denoted as the **HBO** and **LBO** modes, respectively.

A. Channel-Selection

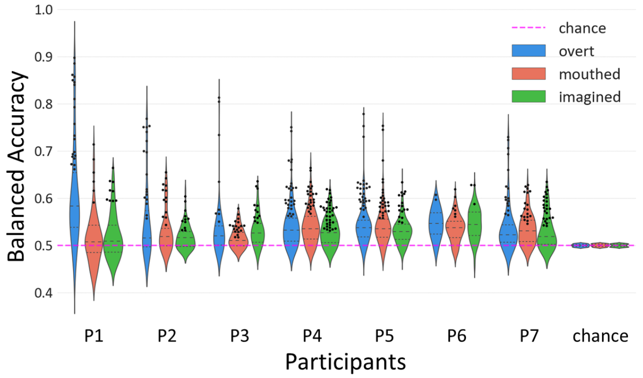


Fig. 3: Distributions of classification performance of decoding models for all channels for the channel selection procedure. The black dots represent the channels with significantly above threshold performance for each group ($p < 0.01$).

Fig. 3 illustrates violin plots of the distributions of averaged balanced accuracy of the 10-fold cross-validation from the channel selection process. To compare to chance-level performance, permutation tests were performed by randomly shuffling the labels and performing the 10-fold non-shuffled cross-validation process 1,000 times. At total of 68 channels across participants were selected for the overt-mouthed pair, 94 for the overt-imagined pair, and 87 for the mouthed-imagined pair.

1) Discrimination of Modes: Fig. 4 illustrates the distribution of average performance of the MD models of selected channels according to HBO relevance for *speech* and *non-speech* conditions. The chance-level is computed as the average of the chance-level distributions extracted from randomly shuffling the labels and training and evaluating the models 1,000 times. The models for all three HBO-LBO mode pairs performed significantly above chance-level ($p < 0.001$) for *speech*, but showed no significant difference for *non-speech* ($p > 0.05$). To validate the channel selection, MD models were generated from the non-selected channels and none were found to perform significantly above chance level ($p > 0.05$).

For the *speech* condition, the overt-imagined models performed significantly better than both the overt-mouthed and mouthed-imagined models ($p < 0.01$), and the overt-mouthed models performed significantly better than the mouthed-imagined models ($p < 0.05$). This is in line with prior studies that showed a greater difference between neural processes involved during overt and imagined, compared to either overt and mouthed or mouthed and imagined, and overt and mouthed compared to mouthed and imagined [8], [13], [14].

While each of the respective models contained at least eight channels per participant, alternately generating models according to LBO relevance (i.e, mouthed-overt, imagined-overt, and imagined-mouthed pairs) yielded fewer than 4 channels in all cases, with four participants having no channels selected. Additionally, no significant difference was observed between the average performances of these channels vs. chance-level ($p > 0.05$). This is in line with evidence from prior studies showing a hierarchy of neu-

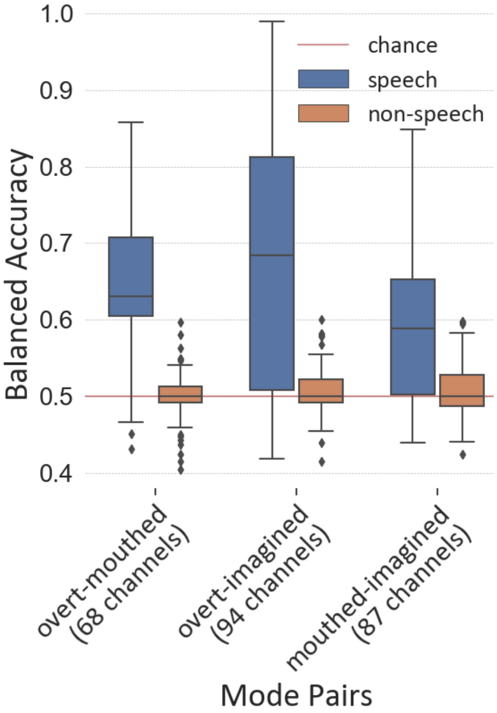


Fig. 4: Box plots of the average balanced accuracy of the selected channels (across all participants) during *speech* and *non-speech* segments.

ral processes with respect to degree of behavioral output, increasing monotonically in imagined, mouthed, and overt speech modes [12], [14].

IV. Conclusion

This study examined models of neural features that discriminate between speech modes that vary with respect to degree of behavioral output. While prior studies have posited common neural substrates underlying these speech modes, the present results further highlight the existence of features that are unique to each mode. Moreover, since the models were separately evaluated for speech and non-speech segments, the results indicated that the differences between these modes are due to speech-related neural processes rather than other factors related to experimental design such as trial ordering or repetition effects. These findings further highlighted the need for careful consideration and treatment when using overt or mouthed speech as surrogates of imagined speech for designing and interpreting imagined-speech decoding models.

Relevant brain regions across modes were found beyond the cortex, bilaterally and at various depths, providing additional evidence of deeper structures' potential relevance in the development of improved speech decoding models. Further work is needed to explore the specific differences of brain regions and networks with respect to speech production across the three modes, as well as whether these results can be generalized across a larger participant pool.

References

[1] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals," *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 140–154, 2011.

[2] C. Herff, D. J. Krusienski, and P. Kubben, "The potential of stereotactic-EEG for brain-computer interfaces: current progress and future directions," *Frontiers in Neuroscience*, vol. 14, p. 123, 2020.

[3] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ECoG using densely connected 3d convolutional neural networks," *Journal of Neural Engineering*, vol. 16, no. 3, p. 036019, 2019.

[4] C. Herff, L. Diener, M. Angrick, E. Mugler, M. C. Tate, M. A. Goldrick, D. J. Krusienski, M. W. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Frontiers in Neuroscience*, vol. 13, p. 1267, 2019.

[5] 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.

[6] 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.

[7] M. Angrick, M. C. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski *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.

[8] M. Perrone-Bertolotti, L. Rapin, J.-P. Lachaux, M. Baciu, and H. Loevenbruck, "What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring," *Behavioural Brain Research*, vol. 261, pp. 220–239, 2014.

[9] C. Cooney, R. Folli, and D. Coyle, "Neurolinguistics research advancing development of a direct-speech brain-computer interface," *IScience*, vol. 8, pp. 103–125, 2018.

[10] N. F. Ramsey, E. Salari, E. J. Aarnoutse, M. J. Vansteensel, M. G. Bleichner, and Z. Freudenburg, "Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids," *Neuroimage*, vol. 180, pp. 301–311, 2018.

[11] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N. E. Crone, J. Rieger, G. Schalk, R. T. Knight, and B. N. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in Neuroengineering*, vol. 7, p. 14, 2014.

[12] K. Okada, W. Matchin, and G. Hickok, "Neural evidence for predictive coding in auditory cortex during speech production," *Psychonomic Bulletin & Review*, vol. 25, no. 1, pp. 423–430, 2018.

[13] W. Zhang, Y. Liu, X. Wang, and X. Tian, "The dynamic and task-dependent representational transformation between the motor and sensory systems during speech production," *Cognitive Neuroscience*, vol. 11, no. 4, pp. 194–204, 2020.

[14] P. Z. Soroush, C. Herff, S. K. Ries, J. J. Shih, T. Schultz, and D. J. Krusienski, "The nested hierarchy of overt, mouthed, and imagined speech activity evident in intracranial recordings," *bioRxiv*, 2022.

[15] P. Roussel, G. Le Godais, F. Bocquelet, M. Palma, J. Hongjie, S. Zhang, A.-L. Giraud, P. Mégevand, K. Miller, J. Gehrig *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.

[16] G. Li, S. Jiang, S. E. Paraskevopoulou, M. Wang, Y. Xu, Z. Wu, L. Chen, D. Zhang, and G. Schalk, "Optimal referencing for stereo-electroencephalographic (sEEG) recordings," *NeuroImage*, vol. 183, pp. 327–335, 2018.

[17] P. Soroush, M. Angrick, J. Shih, T. Schultz, and D. Krusienski, "Speech activity detection from stereotactic EEG," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2021, pp. 3402–3407.

[18] P. Soroush, C. Herff, S. Ries, J. Shih, T. Schultz, and D. Krusienski, "Contributions of stereotactic EEG electrodes in grey and white matter to speech activity detection," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 4789–4792.