Speech Activity Detection from Stereotactic EEG

P. Z. Soroush[1], M. Angrick[2], J. Shih[3], T. Schultz[2], and D. J. Krusienski[1]

2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 978-1-6654-4207-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/SMC52423.2021.9659058

Abstract— **Recent studies have shown promise for designing Brain-Computer Interfaces (BCIs) to restore speech communication for those suffering from neurological injury or disease. Numerous BCIs have been developed to reconstruct different aspects of speech, such as phonemes and words, from brain activity. However, many challenges remain toward the successful reconstruction of continuous speech from brain activity during speech imagery. Here, we investigate the potential of differentiating speech and non-speech using intracranial brain activity in different frequency bands acquired by stereotactic EEG. The results reveal statistically significant information in the alpha and theta bands for detecting voice activity, and that using a combination of multiple frequency bands further improves performance with over 92% accuracy. Furthermore, the model is causal and can be implemented with low-latency for future closed-loop experiments. These preliminary findings show the potential of cross-frequency brain signal features for detecting speech activity to enhance speech decoding and synthesis models.**

I. Introduction

Speech is the foremost modality of human interpersonal communication. Individuals suffering from severe neurological disease or injury can completely lose the ability to speak. Brain-Computer Interfaces (BCIs) have shown promise as assistive technologies for communication and control for the disabled [1], [2]. However, the complex nature of speech makes reconstructing naturalistic speech from brain signals extremely challenging. Invasive recording of brain activity, such as electrocorticography (ECoG) [3] and stereotactic electroencephalography (sEEG) [4], can provide high spatiotemporal resolution and have shown promise for designing BCIs for speech decoding and synthesis [5], [6], [7], [8], [9].

Earlier studies in this area have utilized brain recordings to synthesize different parts of speech, including phonemes, characters, and words [5], [8], [9], [10]. One commonality that these and other studies share is the focus on features extracted from broadband gamma activity. While the significance of the information extracted from broadband gamma in representing different tasks (e.g., speech, hand movement [11]) is undeniable, the information in other frequency bands can also help enhance the performance of the models. A few previous studies have investigated extracting features from multiple frequency bands together for speech-related tasks [12], [13], [14]. While these studies highlight the potential contributions of multiple bands, a comprehensive characterization of the individual and joint band contributions was not provided.

The present study aims to investigate the conventional frequency bands (delta, theta, alpha, beta, low-gamma, and broadband gamma), individually and jointly, in the design of a model to discriminate speech and non-speech from intracranial brain activity. This analysis can provide insights into new potential features for enhancing the performance of speech BCIs.

II. Methodology

A. Participants and Data Collection

sEEG data were collected from 4 native English-speaking participants being monitored as part of treatment for intractable epilepsy at UCSD Health. The locations of sEEG electrodes were determined solely based on the participants' clinical needs. A subset of the implanted electrodes for each participant were determined to be in or adjacent to brain regions associated with speech and language processing. Fig. 1 shows the axial view of the depth electrode locations for each participant.

For the experiment, participants alternated trials of vocalizing and generating inner speech for the same sentences. The Harvard sentences [15] were selected for recitation because they are phonetically-balanced through inclusion of specific phonemes at the same frequency they appear in conversational English. The sentences were displayed on a computer monitor and also simultaneously narrated via computer speakers. Immediately following this presentation, participants were prompted for the trial type using vocalize or inner speech icons, respectively, and the participants were asked to perform the associated task immediately upon presentation of the icon within a 4-second interval before the subsequent trial. Data were collected for 25-50 sentences per participant. The intracranial signals and speech via a microphone were simultaneously recorded during the experiment, digitized at 1,024 Hz and 44,100 Hz, respectively. For this preliminary study, only the data from the vocalized trials were used to extract speech and non-speech sections from the audio recordings.

B. Labeling the Audio Files (Speech vs. non-speech)

The overt-speech sections were manually transcribed using wavesurfer [16]. The transcription was performed for separate analyses but was convenient for labeling the audio files as speech and non-speech. For speech and non-speech labeling, a 10 ms frame with no overlap was shifted across the audio file, and the transcription onset/offset timings were

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

[1] P. Z. Soroush and D. J. Krusienski are with Department of Biomedical Engineering, Virginia Commonwealth University, Richmond, VA, 23219 USA zanganehp@vcu.edu

[2] M. Angrick and T. Schultz are with the Cognitive Systems Lab, University of Bremen, Bremen, Germany

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

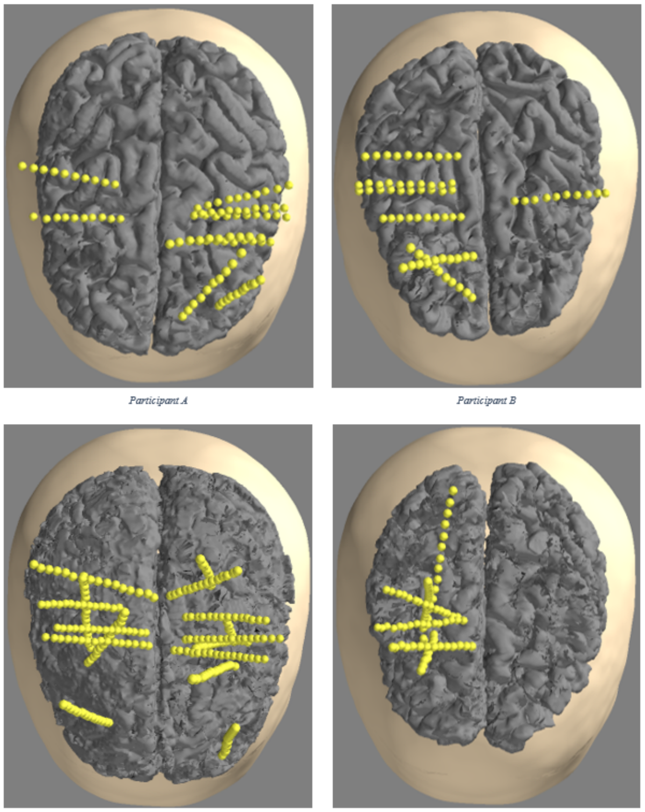


Fig. 1: Axial view of the sEEG depth electrode locations for each of the 4 participants. The frontal lobe is toward the bottom. Note that this represents projections onto the 2D axial plane and the individual electrode shafts have different trajectory angles.

used to create the label. The frame length was chosen to be 10 ms to better represent brain signals' nonstationary nature and the fast changes of speech activity for eventual closed-loop implementation.

Each frame was labeled as *speech* if at least half of the frame length overlapped with a transcribed word and as *non-speech* otherwise. Overall, this yielded 49.6% non-speech frames, indicating that the classes are approximately balanced.

C. Data Preprocessing and Feature Extraction

All sEEG data were visually inspected and noisy or anomalous channels and trials were excluded from the analysis. The resulting sEEG channels were re-referenced using the Laplacian method [17], [18]. The sEEG channels were then normalized by removing the mean and scaled to unit variance. To maintain consistency with a related but separate dataset collected using different equipment, the sEEG signals and the audio signals were resampled to 1,200 Hz and 48,000 Hz, respectively.

The narrow-band amplitude envelope of each normalized sEEG channel was computed in the conventional frequency bands: delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), low-gamma (30-55 Hz), and broadband gamma (65-170 Hz). The subsequent preprocessing was performed for each frequency band.

To compute the amplitude envelopes, using the extracted 10 ms frames from the labeled audio signals, the sEEG channels were segmented over a specified temporal window around this frame. This window was chosen such that it would not extend beyond the 10 ms frame to emulate causal, real-time performance for the model. The window length was chosen to be 310 ms (from 300 ms before the frame's start until the end of the 10 ms frame). However, this is insufficient for the lower frequency bands as at least 3-4 cycles are needed to convey meaningful information in a particular band. Hence, for delta, theta, alpha, and beta bands, the duration of four cycles of the lowest frequency in the band was chosen as the window onset, and the offset was always fixed at the 10 ms frame length (e.g., for 4-8 Hz theta band, the window onset is 1 second (4 cycles x 0.25 s/cycle) before the start of the frame, giving a 1.01 s window length). These longer windows were then segmented to the standard 310 ms window length.

Each window was band-pass filtered over the frequency range of the specific band using zero-phase, Butterworth, sixth order, IIR filters. An additional 118-122 Hz notch filter was applied to broadband gamma to suppress the second harmonic of the line noise at 120 Hz.

The 310 ms filtered windows were divided into thirty-one 10 ms segments. The signal energy of each segment was calculated as the feature. The resulting feature space was #10-ms frames × #channels × #segments for each frequency band, which was concatenated to 2D (#frames × features).

D. Model Training and Evaluation

1) Logistic Regression Model: For each frequency band, a Logistic Regression (LR) model was trained. Additionally, L1 regularization was used, which shrinks the less important feature's coefficients to zero and can provide a convenient interpretation of individual feature contributions based on the non-zero classifier weights. A Proximal Adagrad Optimizer (alpha = 0.05 and L1 = 0.005) with SoftMax function was selected for training the classifier.

A 10-fold cross-validation analysis was employed. To prevent training bias, the nine folds of train data were normalized to zero mean and unit variance, and the same normalization parameters were applied to the test folds. One tenth of the train data was used as a validation set to optimize the hyperparameters of the training models (e.g., alpha and L1 of LR models). Additionally, to establish the chance classification level, a randomization test was performed where the labels of all trials were randomly shuffled and the 10-fold cross-validation process was repeated for 1000 separate randomizations of the labels.

For cross-frequency analysis, a voting scheme was implemented on the single-band classifiers. Since a voting scheme with unbiased weights needs an odd number of inputs to make a decision, the frequency band classifier with the lowest accuracy in the previous step was excluded for each participant. Using a 10-fold cross-validation analysis,

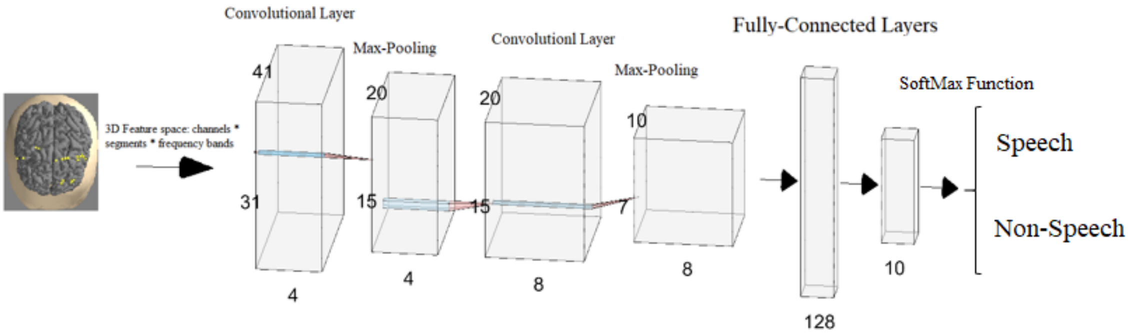


Fig. 2: Architecture of CNN models. The input layer is #channels × #segments × #frequency-bands. Model in the figure depicts a representative model, and the convolutional layers' width change based on the number of channels selected by the LR models.

five separate models were trained using the features of 5 frequency bands. The label with three or more votes was chosen as the label of the frame. The same process was also repeated using the three frequency bands with the highest individual accuracies.

2) Convolutional Neural Network Model: Convolutional neural networks (CNN) have been shown to be promising classification models for different BCI tasks [6], [19], [20], [21]. Therefore, for each participant, a CNN model was implemented for the single and cross-frequency analysis. Due to the small number of classes in the task (2 classes of speech and non-speech), a conventional CNN design (convolutional layers followed by fully-connected layers) was employed with an attempt to lower the computational cost (e.g., number of layers and filters) while maintaining a high accuracy. Fig. 2 depicts the architecture of the final CNN model. Where applicable, a ReLU activation function was implemented. To prevent from overfitting, we added a dropout layer between the last convolutional layer and the first dense layer. An Adam optimizer and categorical crossentropy loss were used for the training of the CNN models.

Using the LR model results, feature selection was performed by excluding the features of each channel that result in an LR weight of zero in all of the features of the channel's segments. The remaining features were converted to 3 dimension (#frames × #channels × #segments). Next, a CNN model was trained using the features and equivalent 10-fold cross-validation. The average of the CNN models' results over all ten folds was computed as the final classification accuracy for each frequency band.

For cross-frequency analysis, features of m frequency bands were used as a multi-layer input which was fed to the CNN model. To be consistent with the LR analysis, for each participant, m was set to 3 and 5, respectively, which represent the number of frequency bands with highest accuracies based on the results of the CNN models from previous section. Since the number of channels selected by the LR models were different for each frequency band, features from all channels were used for these CNN models. For each participant, it was confirmed that sufficient training samples were available for the CNN models to converge.

E. Evaluation of Signal Energy of Different Bands as Feature

In addition to model training and evaluating features using the L1 loss of the LR models, the statistical significance of signal energy in the various bands and channels was evaluated as a feature for speech-non-speech classification. The features of each channel and frequency band were first averaged over speech and non-speech trials separately. Next, a Wilcoxon signed-rank test followed by Bonferroni-Holm correction ($\alpha = 0.05$) was applied to reveal the features exhibiting significant differences between speech and non-speech.

III. Results

Fig. 3 illustrates the results of the 10-fold cross-validation analysis with LR and CNN models, respectively, over all frequency band combinations. On average, the alpha band features provided the best performance across participants and single bands. In contrast, the low-gamma band exhibited comparatively lower performance to the other bands. It is observed that models trained with cross-frequency features perform better than those trained with single frequency features. On average, the CNN models performed slightly better than the LR models, which may be attributed to the CNN models further optimizing the features pre-selected by the LR models.

The chance-level randomization analyses yielded the expected accuracy of approximately 50%. In comparison to the results in Fig. 3, all models performed well above this chance level. However, for participants A and C, the models performed noticeably better than participants B and D, possibly due to the more comprehensive coverage of the left hemisphere and language cortex [22], [23]. This suggests that broad brain areas may contribute to speech vs. non-speech discrimination, but speech-associated areas provide the most impactful contributions.

Fig. 4 shows the speech/non-speech histograms of alpha and broadband gamma energy across trials for a representative dominant channel. As expected, the energy of the broadband gamma increases during speech in comparison to non-speech, while the converse is observed for the alpha band. Furthermore, the relative means and variances of

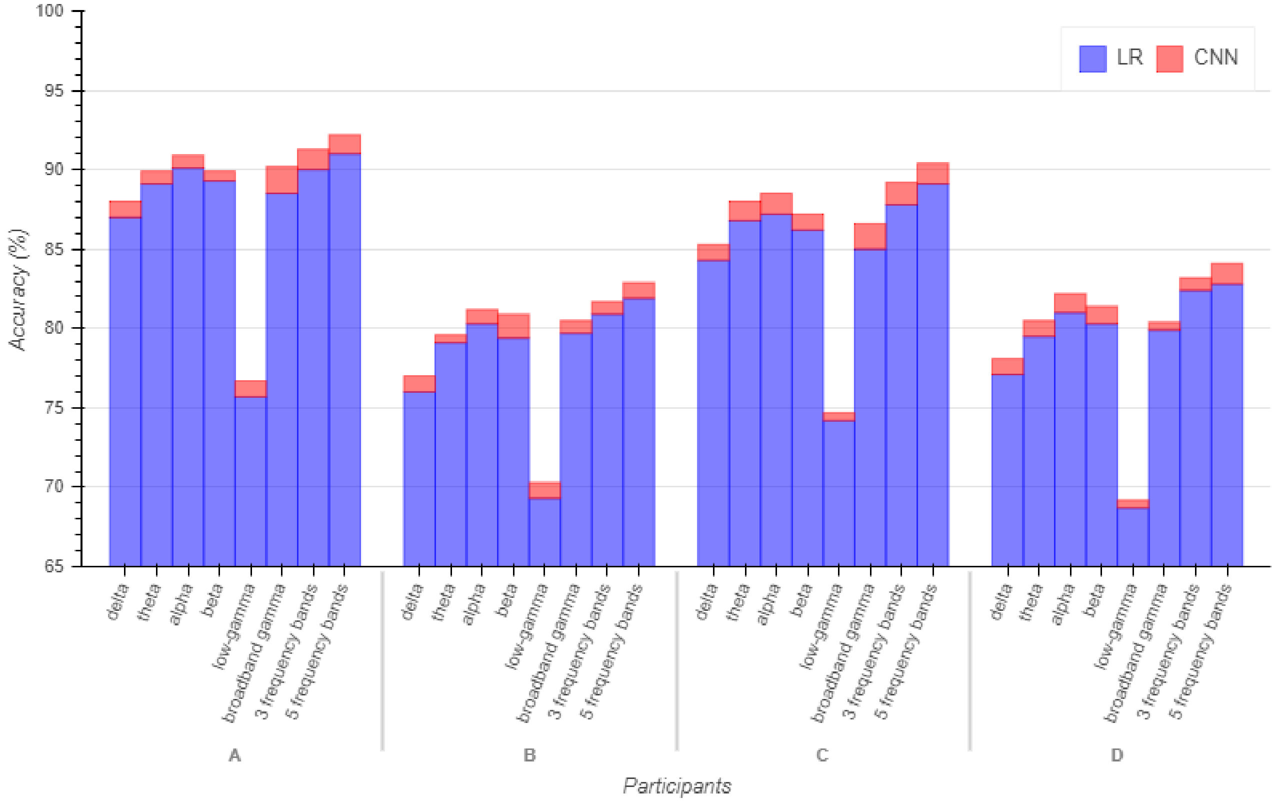


Fig. 3: Classification performance of LR and CNN models for various combinations of frequency band features for each participant (chance level is 50% on average).

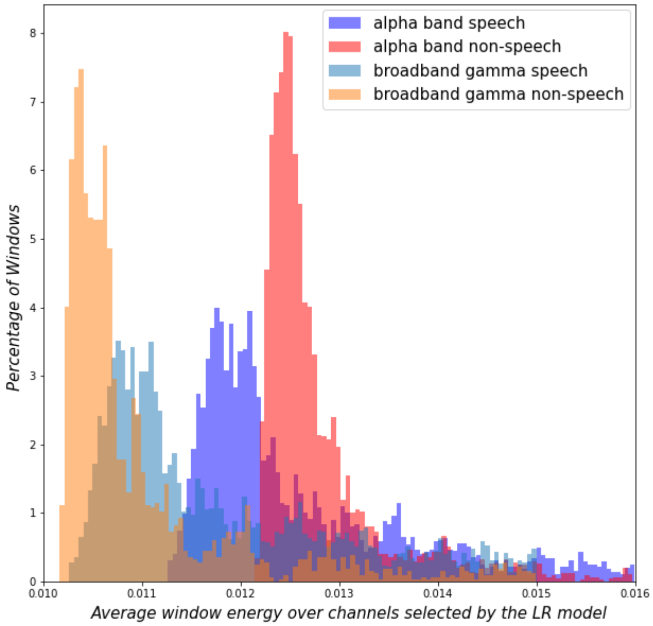


Fig. 4: Speech vs. non-speech histograms of alpha and broadband gamma energy averaged across trials for a representative participant (participant A).

the alpha and broadband gamma activity are comparable, supporting the respective performance results of the LR and CNN models.

To explore the spatio-temporal contributions of the frequency features on the classification, the individual feature weights of the LR model were examined. The weights array was reshaped to #channels × #time samples, and the channels having zero weight in all time samples were excluded. For each frequency band, the average of the ten models trained over the ten folds was computed. Fig. 5 shows the absolute value of this average for a representative participant's (participant A) alpha band's features. It is observed that the dominant channels in the relevant temporal shafts have the largest values close to or on the speech frame (0-1 on the x-axis).

The same procedure was applied to the LR models' weights trained with the same data and randomly shuffled labels. In contrast to Fig. 5 that shows dominant contributions of specific channels in the model, as expected, the randomization test largely yielded a somewhat random distribution of weights across all channels without dominant channels.

The statistical significance of channels between speech and non-speech was explored using a Wilcoxon signed-rank test. For each participant, the significant channels (p-value $\leq$ 0.05) were similar across the frequency bands. Fig. 6 shows the group of statistically significant channels for participant A for the top-performing bands (alpha, beta, theta, and

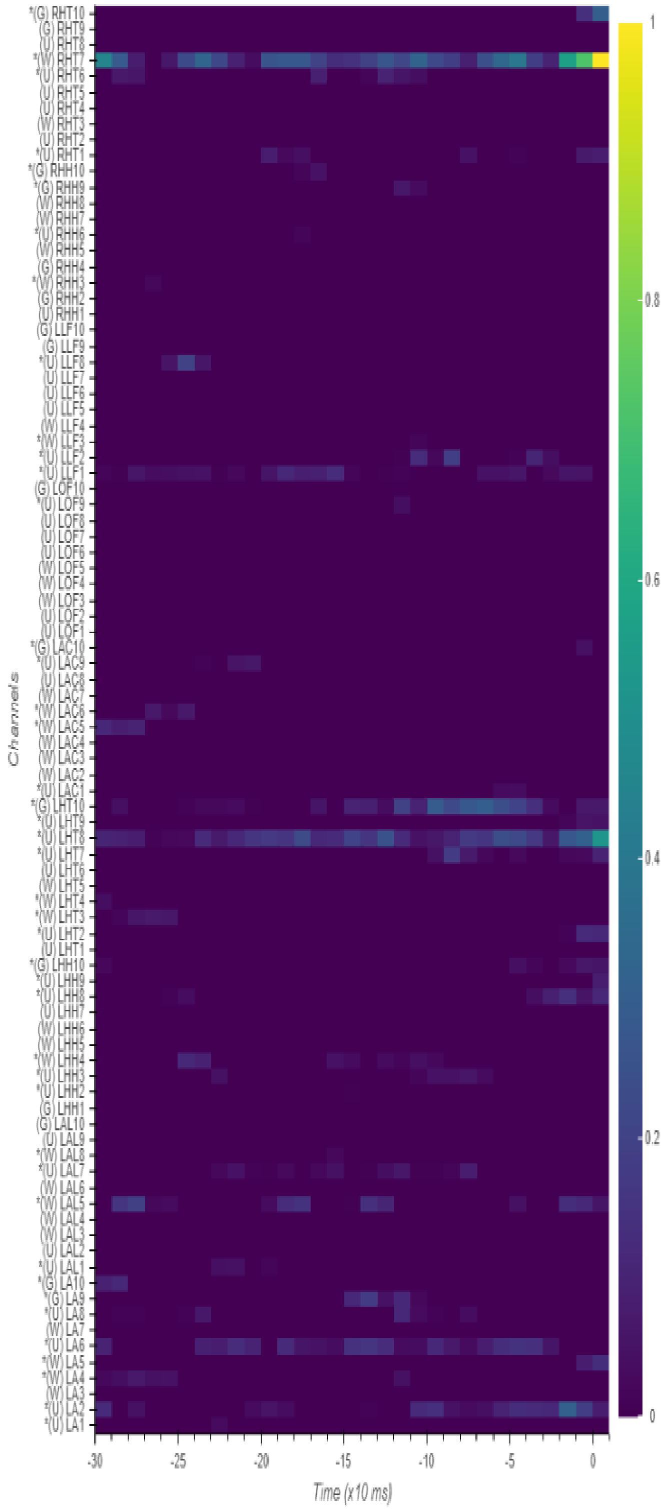
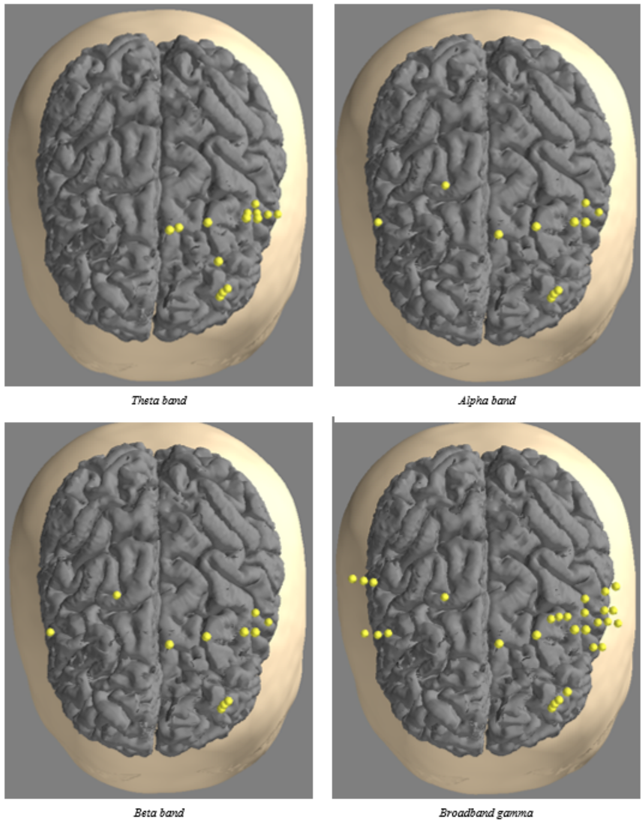


Fig. 5: Absolute value of average LR model weights across 10-folds for a representative participant and frequency band (alpha band's features of participant A). The channel labels are grouped by electrode shaft, with left (L) and right (R) hemisphere designations and numbers 1 and 10 representing depth and superficial contacts, respectively. The letters in parenthesis indicate the location of the electrode on gray (G) matter, white (W) matter, or unknown (U). The channels selected by the model (with nonzero weight) are marked by *.



Theta band

Alpha band

Beta band

Broadband gamma

Fig. 6: Channels with statistically significant activity during speech vs. non-speech for participant A.

broadband gamma).

For participants with coverage of both hemispheres (i.e., participants A and C), the significant channels were located in the left hemisphere, as expected, including those located within proximity to Broca's area. It was also observed that significant channels were located in both grey and white matter, which warrants further investigation.

The channels resulting from the significance tests were compared to Fig. 5 for the LR models. It was found that the majority of statistically significant channels overlapped with the dominant channels from the LR models for each participant and frequency band, further validating the LR model with L1 regularization approach for feature selection.

IV. Conclusion

This study examined two classification models and various combinations of frequency band features from sEEG signals for classification of speech versus non-speech. The results suggest that all examined frequency bands contain significant information to distinguish speech and non-speech well above the chance level. While the broadband gamma features commonly used for intracranial BCI research performed well, it was observed that alpha band activity can achieve comparable or even superior performance to broadband gamma for this task. Moreover, using a combination of bands in a CNN model resulted in a performance of over 92%, compared to the 50% chance level. These findings highlight the potential of exploring spectral features beyond conventional broadband gamma for improving the performance of intracranial speech BCIs. Further work is needed to explore the specific contributions of the superficial and deeper brain areas related to speech/language production and reception on the models, as well as whether these results can be generalized across a larger participant pool.

References

[1] P. Z. Soroush and M. B. Shamsollahi, "A non-user-based BCI application for robot control," in *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, 2018, pp. 36–41.

[2] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "Toward enhanced p300 speller performance," *Journal of neuroscience methods*, vol. 167, no. 1, pp. 15–21, 2008.

[3] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals," *IEEE reviews in biomedical engineering*, vol. 4, pp. 140–154, 2011.

[4] 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.

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

[6] 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.

[7] C. Herff, G. Johnson, L. Diener, J. Shih, D. Krusienski, and T. Schultz, "Towards direct speech synthesis from ecog: A pilot study," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 1540–1543.

[8] P. Sun, G. K. Anumanchipalli, and E. F. Chang, "Brain2char: a deep architecture for decoding text from brain recordings," *Journal of Neural Engineering*, vol. 17, no. 6, p. 066015, 2020.

[9] 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.

[10] E. M. Mugler, M. C. Tate, K. Livescu, J. W. Templer, M. A. Goldrick, and M. W. Slutzky, "Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri," *Journal of Neuroscience*, vol. 38, no. 46, pp. 9803–9813, 2018.

[11] 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.

[12] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "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.

[13] V. G. Kanas, I. Mporas, H. L. Benz, K. N. Sgarbas, A. Bezerianos, and N. E. Crone, "Joint spatial-spectral feature space clustering for speech activity detection from ecog signals," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1241–1250, 2014.

[14] 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. 9, no. 1, pp. 1–12, 2019.

[15] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[16] K. Sjölander and J. Beskow, "Wavesurfer-an open source speech tool," in *Sixth International Conference on Spoken Language Processing*, 2000.

[17] 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.

[18] M. R. Mercier, S. Bickel, P. Megevand, D. M. Groppe, C. E. Schroeder, A. D. Mehta, and F. A. Lado, "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.

[19] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[20] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[21] M. Angrick, C. Herff, G. Johnson, J. Shih, D. Krusienski, and T. Schultz, "Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings," *Neurocomputing*, vol. 342, pp. 145–151, 2019.

[22] J. I. Skipper, S. Goldin-Meadow, H. C. Nusbaum, and S. L. Small, "Speech-associated gestures, broca's area, and the human mirror system," *Brain and language*, vol. 101, no. 3, pp. 260–277, 2007.

[23] C. Wernicke, "The aphasic symptom-complex: a psychological study on an anatomical basis," *Archives of Neurology*, vol. 22, no. 3, pp. 280–282, 1970.