A Novel Idea of Imagine & Real Speech Decoding Model Using Electrocardiogram to Enhance BCI

Chowdhury S. Islam

***Abstract*— Many researchers have tried to decode talked and guessed speech instantly from brain signals towards the growth of a raw-speech BCI. This paper intends to feature extraction and decoding, using the electrocorticoscope (ECoG), the auditory and articulatory features of the motor cortex. Consonants were selected as auditory depictions, and both positions of articulation and manners of articulation were selected as articulatory depictions. The auditory and articulatory representations were decoded at different time lags concerning the speech onset to find optimal temporal decoding parameters. Moreover, this work explores the role of the temporal lobe during speech production directly from ECoG signals. Also, their temporal propagation before and after the speech onset was performed using classification and statistical tests. A novel decoding model using temporal lobe activity was developed to predict a spectral representation of the speech envelope during speech production. Deep learning was utilized in our analysis. This new knowledge may be used to enhance existing speech-based BCI systems, which will offer a more natural communication modality. Also, the work contributes to the field of speech neurophysiology by providing a better understanding of speech processes in the brain.**

Keywords*— *electrocorticogram; EEG; fMRI; deep learning; BCI;

I. Introduction

The main intention of this research is that it specifies the purpose of the temporal lobe through speech construction. Although the role of the temporal lobe is known during speech perception (especially, the auditory cortex), its role during speech production is still not well-defined. Discovering the role of the temporal lobe while speech production takes place can increase the amount of information that is obtained during speech production, which will increase the efficiency of a speech-based BCI system [1]. These analyses highlight the temporal propagation of articulatory and auditory features with respect to the onset at a high temporal resolution. Previous researchers have not used ECoG in such a way [2]. Instead, there were attempts to understand the temporal differences between each representation using fMRI [3] which has a poor temporal resolution. Secondly, these analyses highlight the role of the temporal lobe during speech production as well as the activation/engagement of the temporal lobe along different time lags with respect to speech onset [4]. A very recent fMRI study indicated that there is predictive coding in the auditory cortex during speech production. In our research confirms this conclusion. Lastly, these analyses highlight the usefulness of deep learning [5,6] as an analysis tool in BCI. These collective findings provide important insights toward developing an efficient speech-based BCI system.

II. Temporal Propagation Characterizing Results

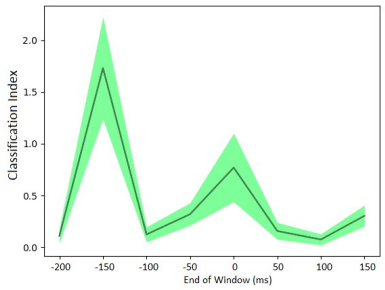
Different sets of words were presented for each subject, and any class that has less than 15 instances is excluded to avoid bias, the number of classes for each subject varies. After excluding rare classes, the number of classes and electrodes for each subject is shown in Table 3.

Table 3 : Subjects classes. M. of A stands for manners of articulation and P. of A stands for places of articulation.

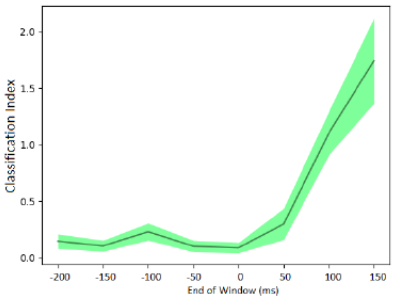
Subject	Consonants classes number	M. of A classes	P.of A classes	Electrodes number
Subject A	11	5	3	31
Subject B	13	5	3	9
Subject C	13	5	3	21

The classification indexes curves are based on 2 different window lengths: 300 ms and 600 ms. The analysis started with a time interval, starting from 500 ms prior to the onset for both 300 and 600 ms windows, and ending at 100 ms or 450 ms after the onset in case of 300 ms and 600 ms windows, respectively. Herein, timings before speech onset will be indicated as negative. The places of articulation classification indexes curve based on a 300 ms window length is depicted in Fig. 1b, which indicates a significant positive gain in the performance in the interval [-.35,.05]s which can be described as the interval where the speech production stage is active. The gradual increase and decrease of the classification indexes curve show the consistency of the temporal propagation for this representation. However, for both consonants and manners of articulation classification indexes in Fig. 1a and Fig. 1c respectively, there are two peaks. The first and the second one occur in the intervals [-450,-150] and [-300, 0] respectively for both representations. This can be interpreted by the window length that was chosen (300 ms), which is too short to capture the information related to these two representations. In order to validate the 300 ms-based analysis as well as to test the effect of the window's length, the analysis was repeated with the 600 ms window's length, which is the same window length of the classification analysis. Since the classification results were statistically significant using this window's length, it is reasonable to speculate a higher classification index, at least in the [-300,300] interval. Fig. 2 depicts that classification indexes were improved as it was speculated. These points to the prolonged-time window of neural activity which is needed to capture the features. The consonants representation's maximum classification index was found in the interval [-200,400] which is depicted in Fig. 2a. For the places of articulation which is depicted in Fig. 2b, the classification indexes reached their maximum in the time intervals [-300,300]. In case of the manners of articulation, the peak of classification index is around [-150,450], where an increase occurs as the window goes to the speech perception

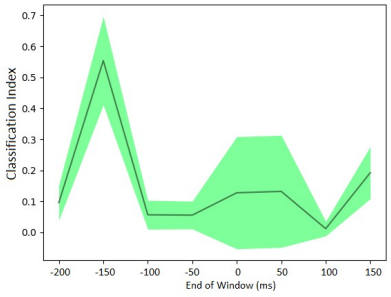
stage. Nevertheless, the values of the classification indexes for manners of articulation are very small compared to the other two representations.



(a) Consonants Classification Indexes

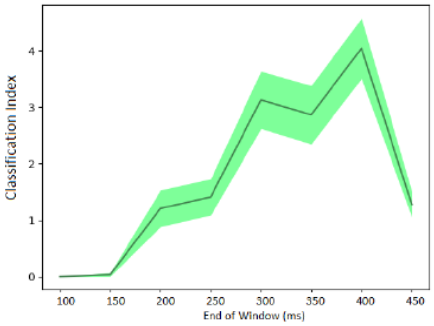


(b) Places of articulation Classification Indexes

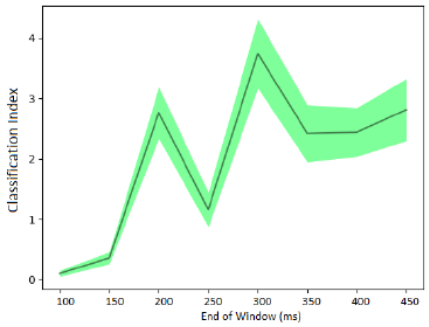


(c) Manners of Articulation Classification Indexes

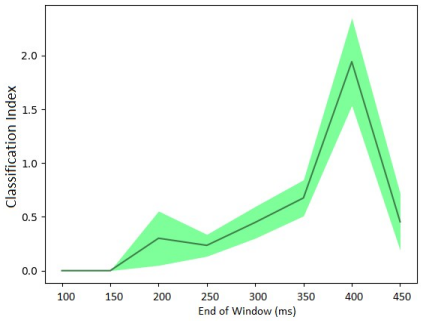
Fig. 1. The estimated mean of the classification indexes for all articulatory and auditory features of subject A using 300 ms window's length. The shaded area represents the 96% confidence interval.



(a) Consonants Classification Indexes

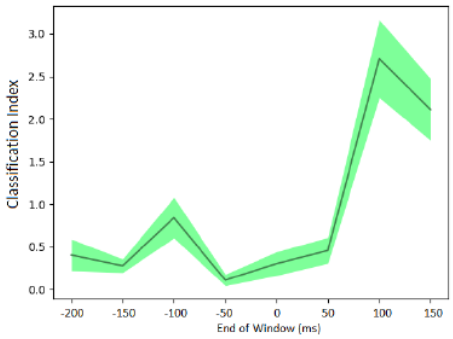


(b) Places of articulation Classification Indexes

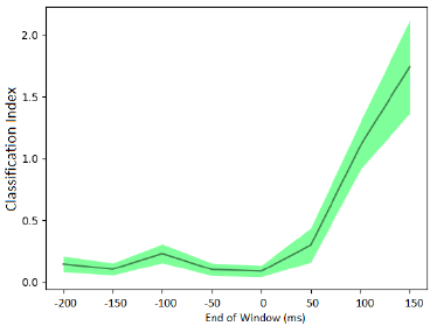


(c) Manners of Articulation Classification Indexes

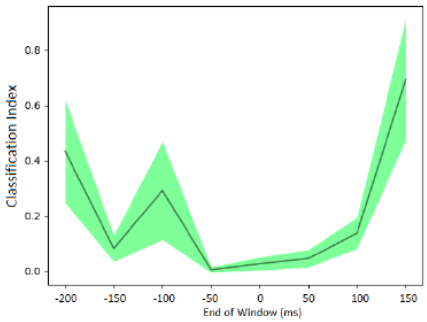
Fig. 2. The estimated mean of the classification indexes for all articulatory and auditory features of subject A using 600 ms window. The area of shaded stands for the 96% assurance interval. The X-axis depicts the shift by 50 ms of the features window. The Y-axis represents the classification index.



(a) Consonants Classification Indexes



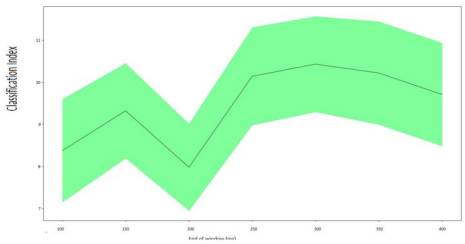
(b) Places of articulation Classification Indexes



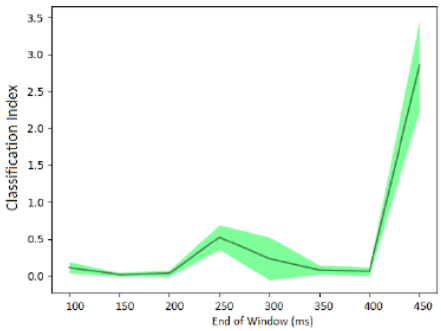
(c) Manners of Articulation Classification Indexes

Fig. 3. The estimated mean of the classification Indexes for all articulatory and auditory features of subject C using 300 ms window. The area of shaded depicts the 96% confidence interval. The X-axis depicts the shift by 50 ms of the features window. The Y-axis represents the classification index

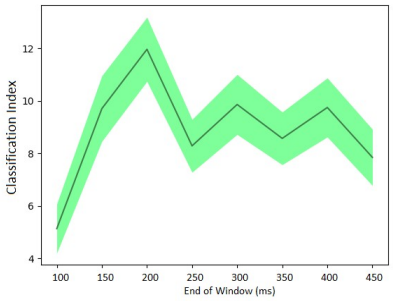
On the other hand, when increasing the window's length to 600 ms, depicted in Fig. 4c, the performance was enhanced and reached its maximum in the interval [-400,200], then decreased and remained constant with a smooth oscillation. The steady state of the curve (smooth oscillations) indicates that no information is gained or lost by shifting the window farther. The places of articulation classification indexes based on the 300 ms window's length are depicted in Fig. 3b, which depicts an increase in the classification indexes all along the interval of [-300,1500] and the maximum classification index occurred in the interval [-150,150]. When increasing the window's length to 600 ms, the places of articulation classification indexes curve did not change as is depicted in Fig. 4b. The consonants classification indexes based on 300 ms window's length is depicted in Fig. 3a. This curve depicts no significant classification index except in the interval [-150,150]. Nonetheless, after increasing the window's length to 600 ms, the consonants classification indexes curve, which is depicted in Fig. 4a, depicts a high classification index in the interval [-300,300] and the performance remains merely constant up to the interval [-100,400], which indicates that no information is gained or lost when shifting the window from [-300,300] to [-100,400]. The analysis of the subject B based on the 300 ms window's length is depicted in Fig. 5. Consonants classification index based on the 300 ms window's length curve in Fig. 5a is maximized during the interval [-150,150] which is the closest interval to speech perception. The manners of articulation classification indexes based on the 300 ms window's length curve is depicted in Fig. 5c, which has a behavior similar to the consonants curve, where it is maximized in the interval [-150,150]. However, the values of



(a) Consonants Classification Indexes



(b) Places of articulation Classification Indexes

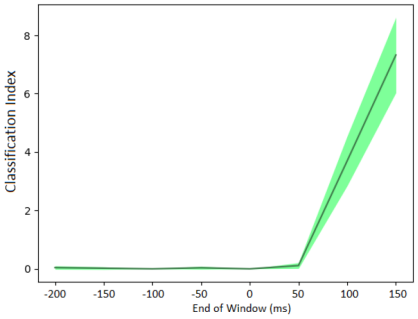


(c) Manners of Articulation Classification Indexes

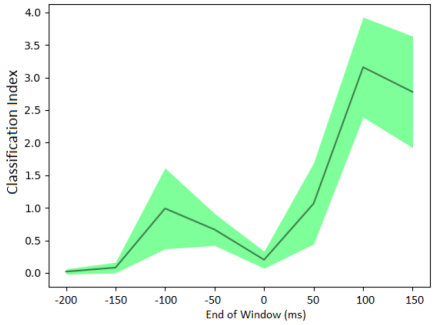
Fig. 4. The estimated mean of the classification indexes for all articulatory and auditory features of subject C using 600 ms window. The area of shaded depicts the 96% interval of confidence. The X-axis depicts the shift by 50 ms of the features window. The Y-axis represents the classification index.

manners of articulation classification indexes are much less than consonants. Places of articulation classification indexes based on the 300 ms window's length curve is depicted in Fig 11b, where it starts to increase in the interval [-400,-100] up to the interval [-200,100] which includes a speech-perception related activity. To further investigate the effect of the speech-perception stage, the analysis was extended to the 600 ms window's length which is depicted in Fig. 6. The manners of articulation classification indexes curve is depicted in Fig. 5c, where it is maximized in the interval [-250,350]. The places of articulation classification indexes curve, which is depicted in Fig. 6b, has a consistent increase starting from the interval of [-500, 100] and then it is maximized in the interval [-.3,.3], and after that, the curve starts to decrease. The consonants classification indexes curve is depicted in Fig. 6a, which has a consistent increase starting from [-500,100] and then it is maximized in the interval [-250,350], then it is followed by a decrease. The decrease of both consonants and places of articulation curves tells that shifting the window farther after the classification index-maximized interval causes loss of information. Based on the extended analysis of each subject, the temporal characterization of the auditory features, which are represented by consonants, based on the 300 ms window length depicts a higher classification index as temporal-parameters of the features go toward the speech perception stage. Furthermore, the temporal characterization based on the 600 ms window length supports this conclusion, that the speech-perception stage is more related to auditory features than speech-production. Neither of the articulatory features

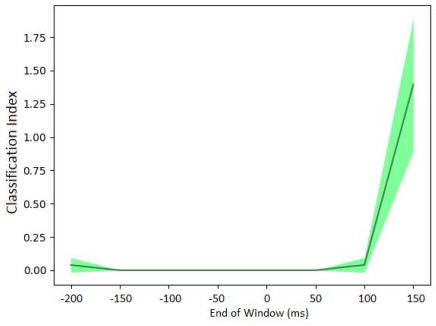
representations, however, depicts consistent temporal classification indexes across subjects for both representations. For instance, subject B showed high manners of articulation temporal classification indexes but also showed weak temporal classification indexes for the places of articulation. In general, the articulatory features and auditory features show higher classification indexes in the case of the 600 ms window's length compared to the 300 ms window's length. In other words, both representations are distributed in a prolonged-time window and they are represented in a time interval larger than 300 ms. Nevertheless, the 600 ms window seems to be too long since the curves usually showed a steady-state classification index when shifting the window farther, for instance, subject C consonants (10c) and manners of articulation (10a) classification indexes curves, in addition to



(a) Consonants Classification Indexes

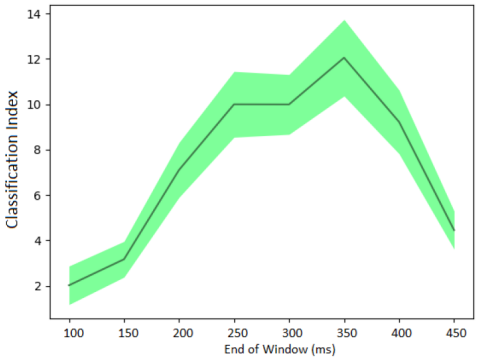


(b) Places of articulation Classification Indexes

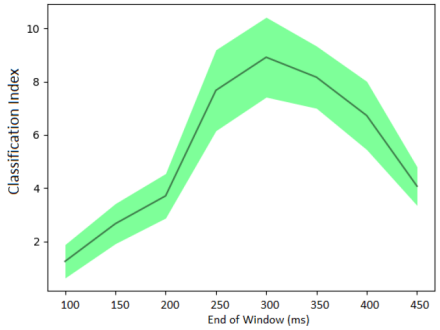


(c) Manners of Articulation Classification Indexes

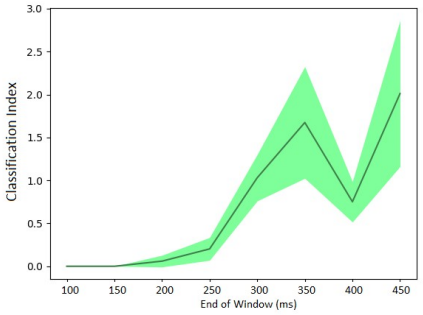
Fig. 5. The estimated mean of the activation Indexes for all articulatory and auditory features of subject B using 300 ms window. The area of shaded depicts the 96% interval of confidence. The X-axis depicts the shift by 50 ms of the features window. The Y-axis represents the activation index.



(a) Consonants Classification Indexes



(b) Places of articulation Classification Indexes



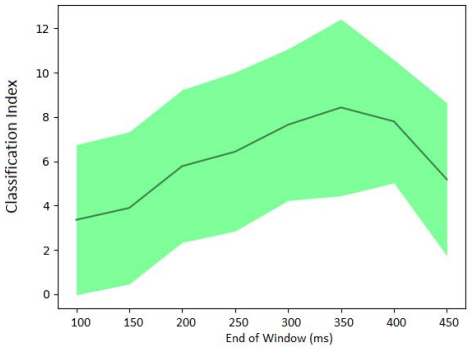
(c) Manners of Articulation Classification Indexes

Fig. 6. The estimated mean of the classification Indexes for all articulatory and auditory features of subject B using 600 ms window. The area of shaded depicts the 96% interval of confidence. The X-axis depicts the shift by 50 ms of the features window. The Y-axis represents the classification index.

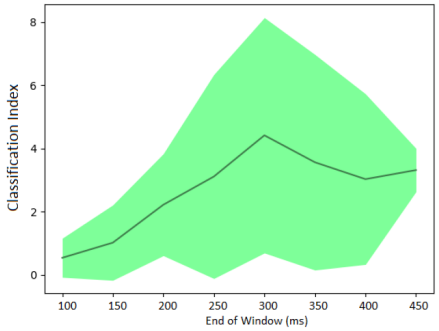
subject A places of articulation classification indexes curve (8b). Further analysis must be done to capture the best window length for each representation.

In general, the temporal characterization of the auditory features has a more consistent classification indexes curve. On the other hand, the temporal characterization of the articulatory features, which are represented by place and manners articulations, varied across subjects. More specifically, subjects A and C have consistent classification indexes for the places of articulation, but subject B has more consistent classification indexes for the manners of articulation. Fig. 7 depicts the typical and standard deviation of all representations across the three subjects. The Consonants curve, depicted in Fig. 7a depicts that the

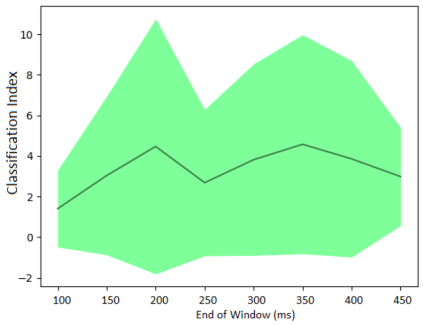
classification indexes increase as the time window is shifted toward speech perception. Places and manners of articulation show a high standard deviation since the classification indexes for these two representations significantly varied across subjects. However, since places of articulation were better represented for two subjects and manners of articulation were better represented for a single subject, the standard deviation of the places of articulation is less than the standard deviation of manners of articulation across subjects.



(a) Average and Standard Deviation of Consonants Classification Indexes across Subjects



(b) Average and Standard Deviation of Places of Articulation Classification Indexes across Subjects



(c) Average and Standard Deviation of Manners of Articulation Classification Indexes across subjects

Fig. 7. The estimated mean of the classification Indexes for all articulatory and auditory features of subject B using 600 ms window. The area of shaded depicts the 96% interval of confidence. The X-axis depicts the shift by 50 ms of the features window. The Y-axis represents the classification index.

III. Modeling Speech-Related Neural Activities in the Temporal Lobe

Most of the selected electrodes of subject A were located at the inferior part of the temporal lobe, which is distant from the auditory cortex. Subject D has the best coverage of the temporal lobe, especially the auditory cortex. The critical value $\alpha_{.05}$ of the chance level Pearson correlation coefficient was estimated in order to have a simple and easy way to read results. The $\alpha_{.05}$ of the chance level Pearson correlation coefficient was calculated for each shift, subject, and frequency group. That is, $4 \times 11 \times 7$ different critical values were obtained, where the 1st number refers to the number of subjects, the 2nd refers to the number of shifts, and the 3rd refers to the number of frequencies group. In order to understand how these critical values differ according to their parameters (i.e., subjects, frequency groups, and lags), they were grouped based on their shift. For instance, all critical values of -500 ms lag were best-fitted to a distribution. It was found that all shift-based groups followed a Normal distribution with very close means [.7,.10]. This means that the distributions of these grouped critical values are merely the same. Therefore, all critical values were bestfitted to a distribution, and it was found that they followed a Gaussian distribution with a mean of .09. This means that different parameters (i.e., subjects, frequency groups, and lags) do not affect the values of $\alpha.05$.

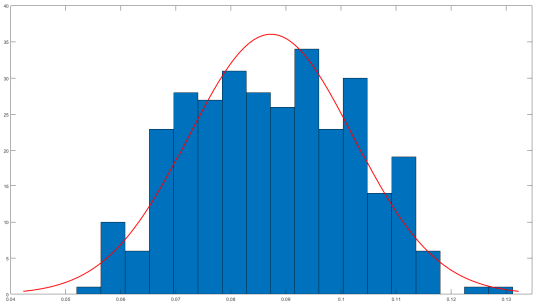


Fig. 8. The critical values obtained from different shifts, frequency groups, and subjects.

This calculation yields $\alpha_{.05} = .0.1156$, which would be considered one of the largest extreme values that a level of chance correlation could take which depicts in Fig. 8.

Fig. 9 depicts the mean, which is represented by the black curve, and the standard deviation, which is represented by the gray shaded area, of the correlation coefficients over different lags across the 4 subjects. The lowest frequency group [1-3]Hz starts from the top, and the higher frequency group [19-21]Hz ends down at the bottom. The x-axis represents the lags starting from -500 ms and up to 500 ms (11 shifts). The red horizontal line represents the critical value $\alpha.05$ of the level of chance. Based on Fig. 9, which depicts the Pearson correlation coefficients from -500 ms to 500 ms lags with 100 ms increase, the propagation curves start to increase from -.5 s to a point in the interval [.1,.2] s and then start to decrease where speech-related activity starts to diminish.

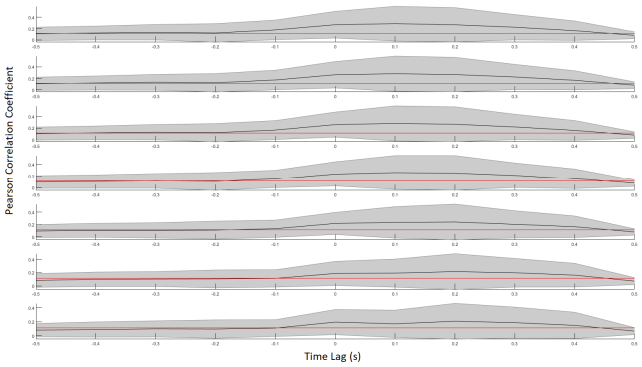


Fig. 9. The Mean and standard deviation of the average correlation coefficient across the 4 subjects for each frequency group between predicted and the actual output of the testing data.

Fig. 10 depicts the Pearson correlation coefficients of subject D. Each curve represents a frequency group. The analysis starts at 500 ms before the onset and ends at 500 ms after the onset with an increase of 100 ms. All values in Matlab on the curve are statistically significant above the level of chance ($p \leq .001$) except when speech leads ECoG signals by 500 ms (i.e., the last point on the curve). For instance, the first group (f_1 in Fig. 1), which is the mean of values at the integer frequencies power in the interval [1-3], has the best correlation with the actual signal and the second higher correlation is assigned to the second group of frequencies [4-6] and so on.

This analysis depicts that there is stronger speech-related neural activity in the very early stage of speech production (-500 ms prior) whereas the previous analysis showed there is either no or very weak speech-related activity in the temporal lobe before -220 ms with respect to the speech onset. It can be interpreted by the LSTM-RNN model is able to capture the nonlinear correlations between the ECoG and speech signals since the prior work was based on Pearson correlation coefficients between speech and neural activities. The performance of the linear regression along with the performance of the LSTM-RNN model is depicted in Fig. 11. For time lags -200 ms prior, the linear model depicts much

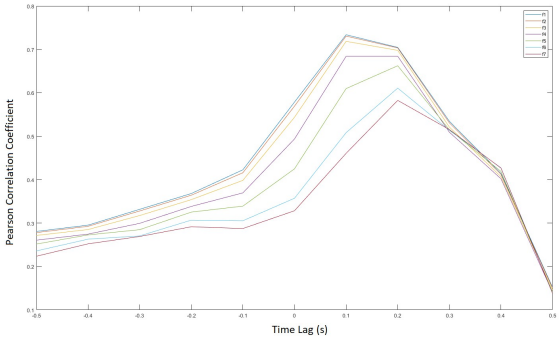


Fig. 10. Decoding the frequencies of speech envelope from subject D temporal lobe using the gamma envelope based on LSTM-RNN model.

weaker correlation coefficients compared with LSTM-RNN model. This may indicate the

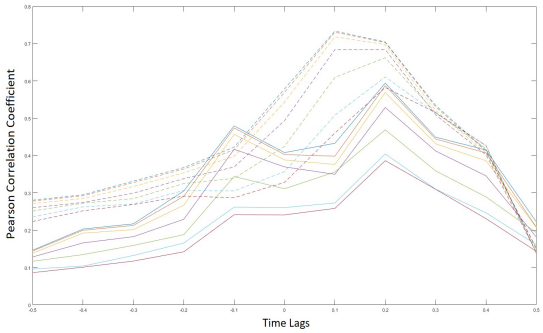


Fig. 11. Linear Model Performance (solid lines) VS LSTM-RNN (dashed lines) performance.

nonlinear relationship between the neural activity in the temporal lobe in the very early the speech production stage and the speech activity. Fig. 11 also depicts the better performance of LSTM-RNN over the linear model [7, 8]. Another interpretation is that, in this work, the spectral power of the speech envelope provides a better representation than the raw envelope used in the prior study.

IV. Conclusion

The two main results of the speech-based BCI indicate that, firstly, the articulatory features appear before the auditory features in the motor cortex by 50 to 150 ms, and auditory features are most relevant to the speech perception stage. Secondly, the temporal lobe is able to predict speech information in the production stage. These two results suggest that multiple decisions can be taken from different regions across different time intervals. Combining these decisions will improve the reliability of the BCI system. For instance, a speech-based BCI can detect the articulatory features from the motor cortex and auditory features from the temporal lobe at the very early stage of speech production. After that, the auditory features are detected from both the motor cortex and the temporal lobe in the late speech production and speech perception stage. Finally, these decisions are combined together to reduce the error and maximize the probability that a detection is correct since more knowledge minimizes the error of machine learning models. This means that a speech-based BCI system can be composed of multiple modules, where each one works on a specific feature representation (e.g., articulatory and auditory) from a specific brain region and at a specific time interval. In other words, each module is specialized in the representation-spatial-temporal decoding technique. Such implementation would lead to improving the real-time speech-based BCI system in a way that if the decision in the early production stage is very confident (i.e., probability of error is too low) then this will help to reduce the response time (i.e., time required to issue a command). This paper contributes to giving possible prototypes of such modules. Another possible usage is that if a speech-based BCI system is mainly implemented to decode the auditory features (e.g., phonemes), an articulatory features-based BCI system can provide support when the former system is confused, in a way that opposes or supports the decision of the auditory features-based system.

REFERENCES

[1] Rashid, M., Sulaiman, N., PP Abdul Majeed, A., Musa, R.M., Ab Nasir, A.F., Bari, B.S. and Khatun, S., 2020. Current Status, Challenges, and Possible Solutions of EEG-Based Brain-Computer Interface: A Comprehensive Review. Frontiers in Neurorobotics.

[2] Okada, K., Matchin, W. and Hickok, G., 2018. Neural evidence for predictive coding in auditory cortex during speech production. Psychonomic bulletin & review, 25(1), pp.423-430.

[3] Berezutskaya, J., Freudenburg, Z.V., Güçlü, U., van Gerven, M.A. and Ramsey, N.F., 2017. Neural tuning to low-level features of speech throughout the perisylvian cortex. Journal of Neuroscience, 37(33), pp.7906-7920.

[4] Conant, D.F., Bouchard, K.E., Leonard, M.K. and Chang, E.F., 2018. Human sensorimotor cortex control of directly measured vocal tract movements during vowel production. Journal of Neuroscience, 38(12), pp.2955-2966.

[5] Riecke, L., Formisano, E., Sorger, B., Başkent, D. and Gaudrain, E., 2018. Neural entrainment to speech modulates speech intelligibility. *Current Biology*, *28*(2), pp.161-169.

[6] Ullah, A., Anwar, S.M., Bilal, M. and Mehmood, R.M., 2020. Classification of Arrhythmia by Using Deep Learning with 2-D ECG Spectral Image Representation. Remote Sensing, 12(10), p.1685.

[7] Peimankar, A. and Puthusserypady, S., 2020. DENS-ECG: A Deep Learning Approach for ECG Signal Delineation. arXiv preprint arXiv:2005.08689.

[8] Zhao, J., Mao, X. and Chen, L., 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47, pp.312-323.