Native language and stimuli signal prediction from EEG

Madhumitha Sakthi
Department of ECE
The University of Texas
Austin, USA
madhumithasakthi.iyer@utexas.edu

Ahmed Tewfik
Department of ECE
The university of Texas
Austin, USA
tewfik@austin.utexas.edu

Bharath Chandrasekaran
Department of Communication Science and Disorders
University of Pittsburgh
Pittsburgh, USA
b.chandra@pitt.edu

***Abstract*—Understanding the neural processing of natural speech processing is an important first step for designing Brain-Computer Interface (BCI) based speech enhancement and speech recognition systems. Complex neural signals like electroencephalography (EEG) are time-varying and has a non-linear relationship with continuous speech. Linear models can decode stimulus features reliably, but the correlation between the reconstructed signal and continuous EEG remain low despite attempts at optimization. In the current application, we demonstrate the utility of a Recurrent Neural Networks (RNN) model to relate various stimuli features such as the envelope, spectrogram to the continuous EEG in a cocktail party scenario. We use a Long Short-Term Memory (LSTM) neural network architecture that has self-connecting loops which help in preserving past information to predict future value. Given that predictability plays a critical role in speech comprehension, we posit that such a neural network architecture yield better results. In attended condition, for native participants, the LSTM models yield 30% and 22% mean correlation improvement and for non-native participants, 43% and 37% improvement over linear models for envelope and spectrogram respectively with EEG. Finally, we have trained a single model to predict the native language of a participant using EEG and it yielded 95% accuracy.**

***Index Terms*—Neural Signal Processing, Speech enhancement, EEG, RNN**

I. Introduction

Developing a Brain-Computer Interface based speech enhancement or speech recognition system for a cocktail party scenario with multiple speakers and noise sources requires a thorough understanding of the neural processing of continuous speech and robust algorithm. A cocktail party scenario is where a person tries to focus on a single speaker among multiple speakers and noise sources. Wherein, the BCI would enhance the speech of the single speaker which would help in better understanding for the listener and provide a robust signal for a speech recognition system. Previous studies have shown that it is possible to perform auditory attention detection i.e., identification of the target speaker, using EEG. For this purpose, the non-invasive EEG signal with a low signal-to-noise ratio is used in the BCI system. The stimuli that is, speech signal is converted to envelope [1] [2] or spectrogram [3] by signal processing. The models are then trained to predict the stimuli signal from EEG. In previous studies, extensive analysis of linear models based on multivariate Temporal Response Function(mTRF) by Crosse, et. al[3] has been done. However, due to the non-linearity of EEG signal, despite best efforts, the correlation between the reconstructed stimuli from the EEG and the actual stimuli remains low. A recent study also used Machine Learning methods[4] with non-linear models to generate the reconstructed envelope signal from EEG. After extensive hyperparameter tuning, they reached seven times higher than the linear baseline in identifying the target speaker. There was also another study where they decoded the target speaker without access to the clean source using Deep Neural Network architecture [6]. This system required long hours of training and access to clean sources during the training phase of the neural networks to identify the target speaker.

In this paper, we are presenting an algorithm for improvement in correlation among the original and reconstructed stimuli using LSTM [7] based RNN architecture and shown that using non-linear models, the correlation improvements are significant and this can be further used in speech enhancement [8] and speech recognition system. Also, we have shown results for native and non-native participants and the algorithm we have developed could be used universally despite the difference in native languages of a participant as long as the language of the stimuli is the same. Finally, we have developed a single model to identify the native language of the participant using their EEG data which could help in Natural Language understanding with a prior knowledge about the native language without explicitly mentioning.

II. MATERIALS AND METHODS

A. Data

There were a total of 15 native English speakers and 14 non-native, Mandarin Chinese speakers in the study. All participants self-reported no previous history of hearing problems or neurological disorders. Each participant provided written, informed consent and received monetary compensation for their participation. The experimental protocol was approved by the Institutional Review Board at The University of Texas at Austin.

Research reported here was supported by the National Institute On Deafness and Communication Disorders of the National Institute of Health under Award Numbers:R01DC015504,R01DC015504, R01DC013315.

The 62-channel EEG data were collected from each participant. The sampling rate of the collected data was 25kHz with the reference as the average of the two mastoid electrodes. They were subjected to two conditions. The *story* and *tone* condition. In the story condition, the participants were asked to pay attention to the story and ignore the tone. In the tone condition, the participants were asked to pay attention to the tone and ignore the story. However, the discrete tone stimulus was meaningful with frequency and duration deviation. They were also asked multiple choice questions to identify if they paid attention subjectively after every session. Each session lasted 1 minute and there were totally 30 sessions for each condition.

B. Stimulus signal processing

The envelope data of the speech stimuli was generated by applying Hilbert transform, then taking the absolute value. Followed by applying a zero-phase shift anti-aliasing filter. The signal was downsampled to 128Hz.

The 16-channel spectrogram data was generated by using band filters between the range of 300Hz to 8000Hz, Short Time Fourier Transform and the resulting signal was downsampled to 128Hz.

The sampling frequency of 128Hz was chosen for the computational efficiency of the algorithm.

C. Linear Model: mTRF

The mTRF based on linear regression with regularization models were developed using the MATLAB toolbox[3]. The ridge parameter for regularization was chosen from 1.0e-1 to 1.0e+5. The time window was chosen from -50ms to 250ms. The EEG signal was downsampled to 128Hz. The stimulus signal was preprocessed to remove ocular artifacts using Independent Component Analysis [9]. Using the mTRF toolbox, leave-one-out cross-validation was performed on the stimuli and EEG for the range of ridge parameter values. The reconstructed stimuli and the actual stimuli were correlated to obtain the Pearson's r correlation coefficient which gives the linear dependency between the two signals. Where Pearson's r correlation is given by equation 1. cov stands for covariance and SD represents standard deviation.

$$r = \frac{Cov(X, Y)}{SD_x SD_y} \tag{1}$$

During training, the model received the EEG related to story and tone. In the testing phase, the model receives the EEG related to story or tone and would predict the corresponding stimuli. In the story condition, i.e., attended condition, if the participant paid attention to the story, predicted stimuli of the story should have a higher correlation with the original stimuli of the story compared to the predicted tone. This establishes a baseline in identifying if the participant was concentrating on the story and hence would result with a lower correlation if the participant paid attention to the tone.

The toolbox returns the correlation value across the ridge parameter values. We have chosen the ridge parameter corresponding to the maximum correlation obtained for the story and tone condition. Hence, the information captured by various ridge parameters are used efficiently. For the spectrogram condition, the model returns correlation across various spectrogram channels and ridge parameter values. Hence, the average across spectrogram channels are taken followed by the maximum correlation among ridge parameters was chosen as the final correlation value.

D. LSTM based Recurrent Neural Networks Model

The RNN model was developed using Keras API, python. The EEG data were normalized. Also, principle component analysis [10] on the EEG data were done in order to decorrelate the EEG channels. Previous studies have shown that the data given to a neural network should be uncorrelated [11]. If two of the inputs are correlated, the neural network weights for e.g. W_1 and W_2 should solve simultaneously to reduce the loss function. Therefore, in this study, we have taken the PCA of individual channels in the EEG data. The train and test split was chosen as 80% and 20% randomly.

The envelope model is a sequential model with a single LSTM layer with 100 cells followed by a dense layer to obtain a single channel output. The loss is calculated as the mean absolute loss [12] with adam [13] optimization. The model was trained up to 1000 epochs with a batch size of 72.

The spectrogram model is also a sequential model with two LSTM layers with 100 cells each followed by a dense layer to obtain the 16-channel output. The loss is calculated as mean absolute loss with adam optimization just like the envelope model. Spectrogram model was trained up to 500 epochs with a batch size of 52. The decreased batch size was to compensate for the increased model complexity due to the additional LSTM layer. This model reconstructs a 16-channel spectrogram data. The channels were individually correlated with the original spectrogram channels and the average of channel correlation was computed to obtain the final correlation.

$$\rho_a = \rho(e_{orig-s}, e_{pred-s}) \tag{2}$$

Equation 2 is the correlation of original envelope of the story with the predicted envelope of the story.

$$\rho_u = \rho(e_{orig-s}, e_{pred-t}) \tag{3}$$

Equation 3 corresponds to the correlation of original envelope of the story with the predicted envelope of the tone. Auditory attention detection [15] [16] is performed when ρ_a is greater than ρ_u. This indicates that the participant paid attention to the story and ignored the tone during the story condition. Also, the participant paid attention to the tone and ignored the story during the tone condition.

E. Model to classify the native language of a participant

Finally, we have developed a single LSTM based RNN model to classify the the native language of a participant. We appended the 14 native and non-native participant's story condition where, they paid attention to the story. We normalized the data and applied principal component analysis to obtain

features that were linearly uncorrelated and retained all the 62 principal components. The model was developed with 3 layers of LSTM, each having 200,100 and 50 units respectively, followed by a dense layer. The train and test split was set to 70% and 30%.

III. RESULTS

The mean correlation is the average of the pearson's r obtained for each class of participants, native and non-native. We have also performed 'Paired Two Samples for Means t-test' [17].

A. Native participants

The Table I shows the correlation for predicted story stimuli with original stimuli in attended condition for the envelope and spectrogram features.

For native participants, the mean correlation for the envelope in mTRF was 0.153 and RNN was 0.199. Hence, the mean correlation improvement of 30%. Similarly, for spectrogram, mTRF was 0.074 and RNN was 0.091, resulting in 22% improvement. The mean correlation is shown in figure 1. From the results, for few participants, the mTRF model performed better than RNN and the reason could be, for these participants, the linear functions in the linear model could fit the data better and has a better performance only these participant. However, the RNN model is generalized and it is able to a give good overall performance for almost all subjects.

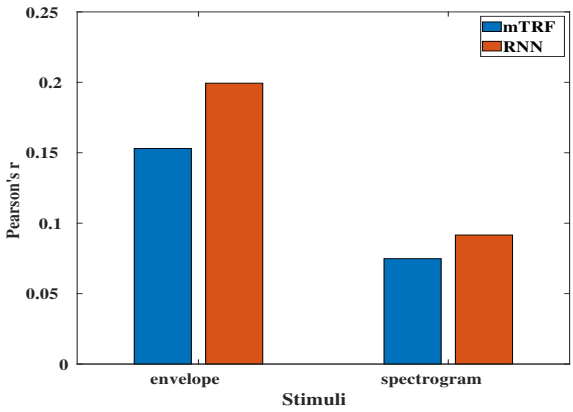

Fig. 1. Native: Mean correlation for predicted and actual stimuli

B. Non-native participants

The table II is the pearson's r correlation value for attended condition in non-native participants. The mean correlation for the envelope in RNN was 0.210 compared to mTRF at 0.147. This resulted in 43% improvement. Similarly, for spectrogram, RNN was 0.103 and mTRF was at 0.075, leading to 37% improvement. The mean correlation is shown in figure 2. Similar to the native case, few participants in non-native has a higher correlation with the linear mTRF model. However, as stated earlier for the native case, the RNN model is generalized and produces better results across participants. Also, overall performance of the RNN yields better correlation with the stimuli compared to linear models.

TABLE I
NATIVE ENVELOPE, SPECTROGRAM CORRELATION

Participants	mTRF envelope	RNN envelope	mTRF Spectro-gram	RNN Spectro-gram
1	0.1664	0.1626	0.0809	0.0699
2	0.1685	0.1761	0.0779	0.0796
3	0.1296	0.2295	0.0561	0.1126
4	0.1378	0.1461	0.0578	0.0596
5	0.2070	0.1160	0.1047	0.0387
6	0.1330	0.1460	0.0621	0.0664
7	0.1149	0.2786	0.0606	0.1468
8	0.1878	0.2439	0.0903	0.1111
9	0.1422	0.2037	0.0711	0.0823
10	0.1630	0.1990	0.0820	0.0978
11	0.1113	0.2292	0.0542	0.1088
12	0.1296	0.2284	0.0696	0.1075
13	0.1415	0.2716	0.0705	0.1513
14	0.1846	0.1671	0.0959	0.0855
15	0.1770	0.1920	0.0873	0.0595

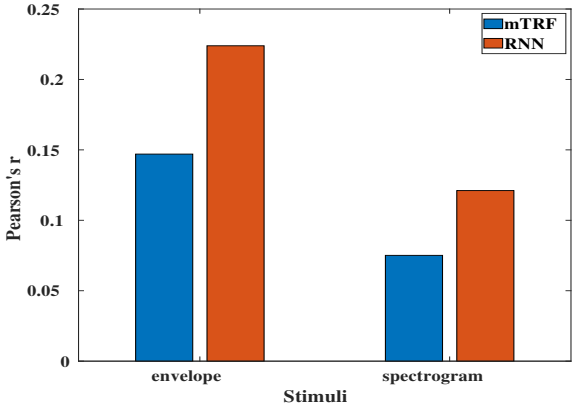

Fig. 2. Non-native: Mean correlation for predicted and actual stimuli

C. Performance analysis of the RNN Models

In the above mentioned neural network architecture, we observed that higher the training epochs, the better the predicted stimuli's correlation. In order to perform a thorough analysis of the network's output, we trained the Envelope model up to 1000 epochs and recorded the correlation for every 100 epochs of training. Similarly, we trained the spectrogram model for 500 epochs and recorded the correlation for every 50 epochs. We observed that for envelope model, even at 1000 epochs the performance was not saturated and we obtained the best results for 1000 epochs. Whereas, in the spectrogram model, the performance started to degrade around 250 epochs indicating that the network was overfitted. This analysis helped in determining the ideal number of training epochs to be chosen for a given model. The change of correlation with the increase in epochs are shown in Figure 4.

TABLE II
NON-NATIVE ENVELOPE, SPECTROGRAM CORRELATION

Participants	mTRF envelope	RNN envelope	mTRF Spectro-gram	RNN Spectro-gram
1	0.1379	0.2033	0.0710	0.0937
2	0.1964	0.1322	0.1017	0.0488
3	0.1315	0.1925	0.0771	0.0969
4	0.1053	0.2776	0.0568	0.1436
5	0.1482	0.2447	0.0787	0.1250
6	0.1177	0.2521	0.0620	0.1317
7	0.1330	0.1530	0.0621	0.0663
8	0.1576	0.2568	0.0740	0.1319
9	0.1671	0.2506	0.0834	0.1318
10	0.0986	0.2244	0.0489	0.1109
11	0.1320	0.2331	0.0633	0.1199
12	0.1991	0.1826	0.0986	0.0799
13	0.1635	0.2088	0.0819	0.1005
14	0.1694	0.1343	0.0915	0.0599

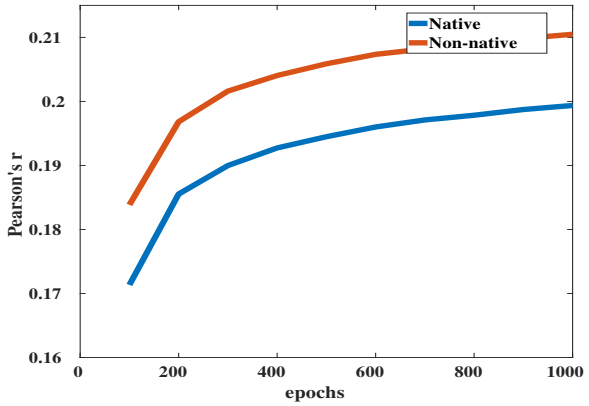

Fig. 3. Envelope Model Analysis

D. Paired Two Samples for Means t-test

We performed 'Paired Two Samples for Means t-test'. The 't-test' is used to compare two averages and determine if their difference is statistically significant. The test was performed on attended conditions across native and non-native participants for envelope and spectrogram features. Due to multiple hypothesis tests on the same data, the Bonferroni correction was made by setting the alpha value to 0.025. All of them exhibited significant score difference except for the native spectrogram case. We believe this is the case for native spectrogram case since the mean difference was the least. However, for envelope preprocessing with the same dataset, the results are statistically significant indicating that spectrogram preprocessing with native subjects need more data.

In native envelope case, RNN (M=0.199, SD=0.002) and mTRF (M=0.153,SD=0.0008); t(14)=2.68,p<0.01.

In native spectrogram case, RNN (M=0.091, SD=0.001) and mTRF (M=0.074,SD=0.0002); t(14)=1.57,p<0.06.

In non-native envelope case, RNN (M=0.21,SD=0.14) and mTRF (M=0.147,SD=0.0009); t(13)=3.5 , p<0.01.

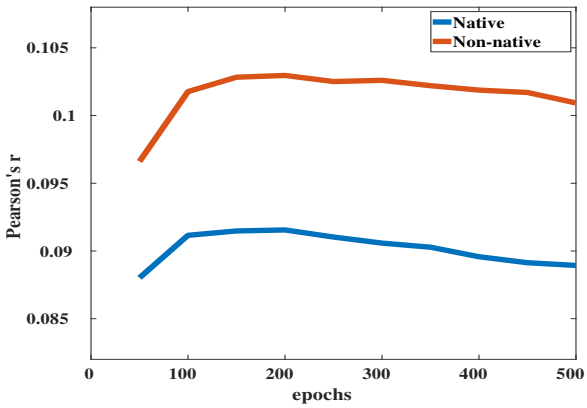

Fig. 4. Spectrogram Model Analysis

In non-native spectrogram case, RNN (M=0.102, SD=0.0009) and mTRF(M=0.075,SD=0.0002); t(13)=2.53, p<0.01.

E. Classification of native language of a participant

The network was trained for 50 epochs where, the training accuracy converged at 96%, with a binary cross-entropy loss of 0.0582. The test accuracy was 95% after 50 epochs.

IV. CONCLUSION

Using Long Short Term Based Recurrent Neural Networks, the mean correlation between the predicted envelope with the actual envelope has significantly improved for the attended condition. For native participants, the overall mean correlation improvement for all participants was 30% and 22% respectively. Also, for non-native participants, the overall mean correlation improvement for all participants was 43% and 37%. This suggests that non-linear RNN models are capable of capturing the information in EEG in order to reconstruct the stimuli signal. Even though the Recurrent Neural Networks model is complex, for a speech enhancement or speech recognition application, it is important to decode the envelope with more precision from the EEG data. This study suggests that for Brain-Computer Interface applications, it would be better to use a Recurrent Neural Networks based model to process EEG data and hence develop a speech enhancement or speech recognition system. Also, the binary classifier based on LSTM RNN yielded 96% train accuracy and 95% test accuracy in identifying the native language of the participants from their EEG data. For the later, as a further study, MFCC features could be used instead of the envelope and we could try and directly decode the speech from the EEG data to develop an unspoken speech recognition system.

ACKNOWLEDGMENT

The data was collected by Rachel Reetzke from the Sound Brain Lab at The University of Texas at Austin.

REFERENCES

[1] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. Speech recognition with primarily temporal cues. Science 270, 303304 (1995).

[2] S. Van Eyndhoven, T. Francart, A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses", IEEE Trans. Biomed. Eng..

[3] Smith, Julius O. Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications, Second Edition,W3K Publishing, http://books.w3k.org/,2007, ISBN 978-0-9745607-4-8.

[4] Crosse M J, Di Liberto G M, Bednar A and Lalor E C 2016 The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli Frontiers Hum. Neurosci.10604

[5] Taillez, T., Kollmeier, B. & Meyer, B. T. (2017), Machine learning for decoding listeners attention from electroencephalography evoked by continuous speech, European Journal of Neuroscience .

[6] J. O'Sullivan, Z. Chen, J. Herrero, G.M. McKhann, S.A. Sheth, A.D. Mehta, N. Mesgarani, "Neural decoding of attentional selection in multi-speaker environments without access to clean sources", Journal of Neural Engineering, vol. 14, no. 5, 2017.

[7] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural

[8] N. Das, S. Van Eyndhoven, T. Francart, A. Bertrand, "EEG-based attention-driven speech enhancement for noisy speech mixtures using N-fold multi-channel Wiener filters", Proc. European Signal Processing Conference (EUSIPCO), Aug. 2017.

[9] Hyvarinen, A. and Oja, E. Independent component analysis: Algorithms and applications. Neural Netw. 13 computation, 9(8):17351780, 1997.

[10] Jolliffe I. (2011) Principal Component Analysis. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg

[11] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Muller. Efficient backprop.In Neural Networks: Tricks of the Trade, pages 950. Springer, 1998

[12] Willmott CJ, Matsuura K. 2005. Advantages of the mean absolute error(MAE) over the root mean square error (RMSE) in assessing averagemodel performance. Climate Research 30: 7982.

[13] Kingma, Diederik P and Ba, Jimmy Lei. Adam: A method for stochastic optimization. arXivpreprint arXiv:1412.6980, 2014.

[14] Cover, T. (2006). Elements of Information Theory. Wiley-Interscience

[15] W. Biesmans et al., Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario, IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. PP, no. 99, pp. 11, 2016.

[16] A. Aroudi et al., Auditory attention decoding with EEG recordings using noisy acoustic reference signals, in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016.

[17] BARTZ, A. E. Basic Statistical Concepts. Minneapolis: Burgess, 198 1, p. 246.