INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Neural Entrainment to Natural Speech Envelope Based on Subject Aligned EEG Signals

Di Zhou[1], *Gaoyan Zhang*[2], *Jianwu Dang*[1,2], *Shuang Wu*[2], *Zhuo Zhang*[2]

[1]Japan Advanced Institute of Science and Technology, Japan
[2]College of Intelligence and Computing, Tianjin key Laboratory of Cognitive Computing and Application, Tianjin University, China

zhanggaoyan@tju.edu.cn, jdang@jaist.ac.jp

Abstract

Reconstruction of speech envelope from neural signal is a general way to study neural entrainment, which helps to understand the neural mechanism underlying speech processing. Previous neural entrainment studies were mainly based on single-trial neural activities, and the reconstruction accuracy of speech envelope is not high enough, probably due to the interferences from diverse noises such as breath and heartbeat. Considering that such noises independently emerge in the consistent neural processing of the subjects responding to the same speech stimulus, we proposed a method to align and average electroencephalograph (EEG) signals of the subjects for the same stimuli to reduce the noises of neural signals. Pearson correlation of constructed speech envelops with the original ones showed a great improvement comparing to the single-trial based method. Our study improved the correlation coefficient in delta band from around 0.25 to 0.5, where 0.25 was obtained in previous leading studies based on single-trial. The speech tracking phenomenon not only occurred in the commonly reported delta and theta band, but also occurred in the gamma band of EEG. Moreover, the reconstruction accuracy for regular speech was higher than that for the time-reversed speech, suggesting that neural entrainment to natural speech envelope reflects speech semantics.

Index Terms: speech envelope, neural entrainment, subject aligned EEG, reconstruction accuracy

1. Introduction

Speech perception, which links auditory and cognitive processes, is the acquisition of communicative information from speech sounds [1]. In current years, studies have extended to investigate how neural activity tracks the acoustic or linguistic information of a continuous speech stream, which is called neural entrainment to the speech signal [2–5]. They found that the neural response in the delta and theta frequency bands could track the speech envelope when listening to speech [6]. In the studies of neural entrainment, the stimuli are always presented to subjects only once to avoid a priming effect. Because it is impossible to specify the event related potentials (ERP) by averaging more trials, this kind of studies employed a system modeling frame to estimate the temporal response functions (TRFs) of the neural system. If we treat the neural system as a linear system, speech signal such as its envelope of the stimuli can be reconstructed from EEG signals, and the reconstruction accuracy is generally used to evaluate neural entrainment. However, the reconstruction accuracy was not so high enough currently. Meanwhile, some contrary results were reported. For example, speech envelope is usually considered related to low-level acoustic feature such as syllable boundary [5], while some studies provided that neural entrainment to speech was stronger when speech was easy to understand [7, 8]. They argued whether or not speech envelope tracking is modulated by high-level language processing. Some other studies defeat that there was no difference in the neural entrainments between accessible and inaccessible speech [9–11]. Here, we speculate that the contrary may be caused by the unexpected noises during collecting EEG data. As well known, the scalp EEG signal is easily contaminated by external noise, such as eye movement, body activity, heartbeat and breath. Single-trial based result is largely dependent on the noise level of EEG signal. If the noises of EEG signal can be largely reduced in this situation, higher reconstructed accuracy of speech envelope and better results can be expected.

In this study, we proposed a method to reduce the external noise for continuous natural speech by aligning the subjects' neural signals. For the same speech stimulus, each subject probably uses the same neural mechanism to process speech, which means the TRFs is similar. In contrast, breath, heartbeat and other external noises from different individuals often occur randomly. Therefore, if we can align the neural responses to corresponding stimulus from multiple subjects and then obtain the average value, such random noises will be reduced by the averaging processing on the EEG signal. As a result, a higher reconstruction accuracy of envelope and a more accurate TRF would be obtained from the averaged subject aligned EEG data than single-trial ones.

In section 2, we will introduce the experiment design and neural entrainment modeling methods in details. Our results will be reported in section 3 and discussed in section 4. In the end, conclusions are given in section 5.

2. Materials and Methods

2.1. Experimental design

2.1.1. Participants

Twenty-two healthy Mandarin Chinese speakers (mean ± standard deviation age, 22 ± 2.4 years; nine men; right-handed) were recruited from Tianjin University and Tianjin University of Finance and Economics. The experiments were conducted in accordance with the Declaration of Helsinki [12] and approved by the local ethics committee. The subjects signed informed consent forms before the experiment and were paid for their participation afterward. All the subjects reported no history of hearing impairment or neurological disorders.

http://dx.doi.org/10.21437/Interspeech.2020-1558

2.1.2. *Stimuli and procedure*

Subjects undertook 48 non-repetitive trials separated into two groups; each trial was around 60 s. One group of the trials consists of 24, short stories with a complete storyline recorded by a male Chinese announcer in a soundproof room. And another group included the left 24 trials, which were the story segments, but played in time-reverse, and was used as a contrast to evaluate whether neural entrainment to speech envelope reflects speech intelligibility. All stimuli were mono speech with 44.1 kHz sampling rate, and the stimulus amplitudes were normalized to have the same root mean square (RMS) intensity. The 48 trials were randomly presented to the subjects. All speech segments were also modified to truncate the silence gaps to less than 0.5s [3].

The experiment was carried out in an electronically and magnetically shielded soundproof room. In the experiment, speech sounds were presented to subjects through Etymotic Research ER-2 insert earphones (Etymotic Research, Elk Grove Village, IL, USA) at a suitable volume (around 65 dB). During each trial, subjects were instructed to focus on a crosshair mark in the center of the screen to minimize blinking, head movements, and other bodily movements. There was a five-second interval between each trial, and the subjects were given a five-minute break every ten trials. After each story trial, subjects were asked immediately to answer multiple-choice questions about the content of the story to ensure that they focused on the auditory task. For the time-reverse trials, we embedded unique tones in some trials to draw more of the subjects' attention to the stimuli. Subjects were requested to detect the tones and indicate how many times they appeared after the trial. The EEG data corresponding to the embedded tones were removed in further analysis.

2.2. Method details

2.2.1. *Data acquisition and pre-processing*

The scalp EEG signal was recorded with a 128-channel Neuroscan Synamps system (Neuroscan, USA) at a sampling rate of 1000 Hz. The electrodes were placed according to the standard 10-5 system, and six channels were used for recording a vertical electrooculogram (VEOG), a horizontal electrooculogram (HEOG), and two mastoid signals. The impedance of each electrode was kept below 5 kΩ during data acquisition. Three subjects' data were discarded in further analysis because they did not give a proper answer for the multiple-choice questions or the electrodes detached during the EEG data recording.

The raw EEG data were pre-processed using the EEGLAB toolbox [13] in MATLAB (MathWorks). This involved removing sinusoidal (i.e., line) noise and bad channels (i.e., low-frequency drifts, noisy channels, short-time bursts) and repairing the data segments. Then, the EEG data was bandpass filtered in the delta band (1-3 Hz), theta band (4-8 Hz), alpha band (9-12 Hz) and beta band (13-30 Hz), and then downsampled to 64 Hz [2, 14]. For the gamma band, the processed raw EEG data were filtered with bandpass of 31-40 Hz and 40-70 Hz, then downsampled to 100 Hz and 150 Hz.

The broadband temporal speech envelopes were obtained from Hilbert transforms [15]. For the following modeling approach, the envelope was then decimated to the same sampling rate as EEG, enabling us to relate their dynamics to the EEG signals.

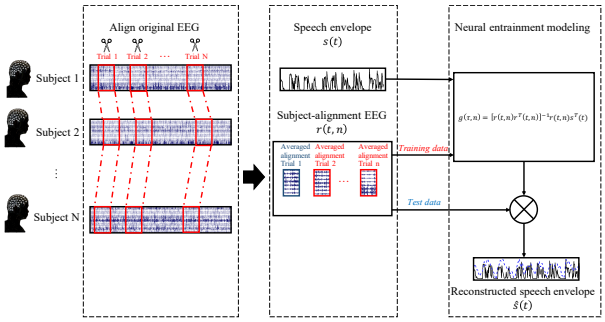

Figure 1: *Subject-alignment based neural entrainment modeling procedure.*

2.3. Subject-alignment of EEG signals

During the experiment, we marked the time trigger for the EEG signal according to the stimuli onset and offset. In the offline analysis, the 48 data epochs (24 story trials and 24 time-reverse trials) were extracted on the basis of the time trigger for each subject. Then, we could separately get data epochs (19 subjects × 48 trials) for story and time-reverse stimuli. We assume that all of the subjects use the same neural mechanism to process the stimulus speech so their TRFs are nearly the same. The averaged alignment data on all subjects is expected to reduce the noises which may be caused by breathing, inattentiveness, etc., through averaging processing. After averaging the subject aligned EEG, we got 48 EEG data epochs.

2.4. Neural entrainment modeling

In this study, we used an mTRF toolbox (https://github.com/mickcrosse/mTRF-Toolbox) to linearly map the speech envelope and the neural response [16]. The main principle is to treat the brain as a linear time-invariant (LTI) system where the output (neural response) of the system is the convolution of the input and a TRF of the brain. The TRF can be considered a filter that linearly transfers the continuous speech envelope to the dynamic neural response. The TRF of the channel n is a function of $\omega(t, n)$ of time t and the output of the neural system is $r(t, n)$ for the same channel n. For an input speech stimulus $s(t)$, the output can be described as:

$$r(t,n) = \sum_{\tau} \omega(\tau, n) s(t-\tau). \tag{1}$$

In a hypothetical LTI system, a backward decoding approach can be modeled using a decoder $g(t, n)$, which is the inverse function of $\omega(t, n)$. Thus, the input speech stimulus $s(t)$ can be reconstructed by filtering the neural response $r(t, n)$ using the decoder function $g(t, n)$. This can be expressed as:

$$\hat{s}(t) = \sum_{n} \sum_{\tau} g(\tau, n) r(t-\tau, n). \tag{2}$$

Where $\hat{s}(t)$ is the reconstructed speech stimuli. Here, the solution of $g(t, n)$ is:

$$g(t,n) = [r(t,n)r^T(t,n)]^{-1} r(t,n) s^T(t) \tag{3}$$

The optimal decoder $g(t, n)$ is acquired by minimizing the mean-squared error (MSE) between the original and reconstructed speech stimuli. The proposed subject-alignment based neural entrainment model is shown in Figure 1.

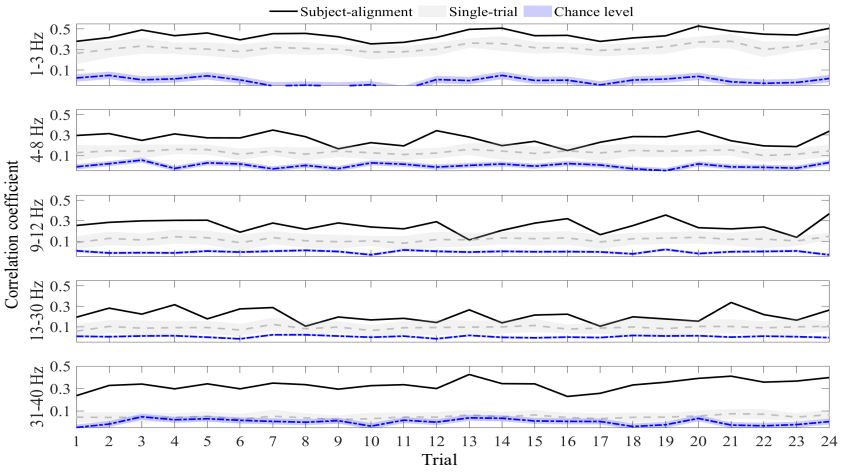

Figure 2: *Comparison of reconstruction accuracies between the proposed subject-alignment based method and the single-trial based method.*

To evaluate our method, we used the subject aligned EEG to reconstruct speech envelope, and compared the reconstruction accuracy of the proposed subject-alignment method with those of the previous single-trial based method. For all training processes, we used a leave-one-out cross-validation procedure, where 23 trials were used for training, and the remaining one trial was used for testing in each fold. Because the parameters of $g(t, n)$ was different in each trial, we used the averaged parameters of the decoder $g(t, n)$ trained on the other 23 trials [17]. In single-trial based method, the decoders were trained based on subject's neural data. Since each subject took part in 24 trials, the procedure was repeated 24 times for each subject. Our proposed method was trained based on the averaged subject aligned data, which was repeated for 24 iterations for the averaged data.

3. Results

3.1. Behavioral results

During the experiment, speech comprehension was evaluated by subjects. For story speech and time-reversed speech, speech comprehension acquired 4.74 ± 0.45 and 1.46 ± 0.81 of the 5 scores respectively (the scores of 5 is very easy to understand and 1 is completely incomprehensible). It means the comprehension of time-reversed speech is very low. For the accuracy of multiple choice questions after each trial, the accuracy of the answers was 88.25 ± 4.62%, indicating that most of the subjects concentrated on the listening task during the experiment.

3.2. Comparison of reconstruction accuracies between the proposed method and single-trial based method

The reconstruction accuracy was evaluated by measuring the Pearson correlation coefficient between the reconstructed speech envelope and the original one. In our study, the efficiency of the proposed subject-alignment method was represented by the correlation coefficient. To quantitatively compare the two methods, the correlation coefficient was firstly transformed into a z value by Fisher's z transformation to satisfy a normal distribution [18]. Then, an analysis-of-variance (ANOVA) of the z values with factors of frequency (different frequency bands) and reconstruction methods (proposed method and single-trial based method) revealed a significant effects on both frequency ($F = 1122.8, p < 0.001$) and reconstruction methods ($F = 623.54, p < 0.001$). The results of ANOVA demonstrate that the reconstruction accuracy of speech envelope of our subject-alignment method is significantly higher than that of single-trial based method. Figure 2 displays the comparisons of reconstruction accuracies using the proposed and the single-trial methods in delta, theta, alpha, beta and low gamma bands. We used a permutation test to compare the predicted accuracy and the chance level and found that our prediction value is 288 times larger than that of the chance level ($p < 0.05$). One can see that more than 5% of the reconstruction accuracy begins to show lower than chance level in 30-40 Hz for the single-trial based method, while our method shows significantly higher than chance level across time in this low gamma band. The single-trial based reconstruction accuracy was not significantly different with chance level in 40-70 Hz, consistent with the literature [6, 7]. Therefore, the results are restricted to 1-40 Hz in Figure 2.

3.3. Reconstruction accuracies for story and time-reversed speeches based on the proposed method

Here, the r_{story} and $r_{time-reverse}$ refer to the averaged correlation coefficients for story trial and time-reversed trial, respectively. The chance level was also acquired by mismatching the neural responses with stimuli data. According to the calculation, the reconstruction accuracy was significantly higher than chance level in all of these frequency bands ($p < 0.05$). The detailed reconstruction results is shown in Figure 3. Figure 4 shows some examples of the reconstructed envelopes obtained in our study. Our results show that neural entrainment to speech also occurs in gamma band, which is less reported in previous research. To clarify whether the reconstruction-accuracy of story is higher than time-reverse speech or not, the values of correlation coefficients were also converted to z values using Fisher's z transformation to satisfy normal distribution. An ANOVA test of z values with main factors of intelligibility (story and time-reversed speech) and frequency bands shows a significant main effect of intelligibility ($F = 78.02, p < 0.001$), indicating that the neural entrainment to intelligible speech is stronger than to unintelligible speech in all frequency bands ($F = 127.52, p < 0.001$).

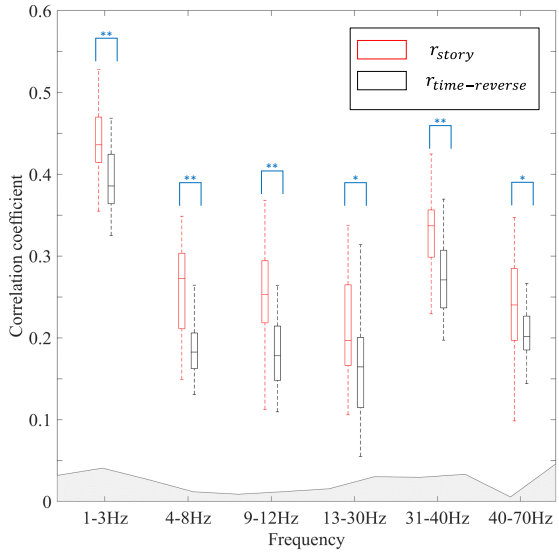

Figure 3: *Correlation between the reconstructed speech envelope and the original speech envelope based on subject-alignment method in story and time-reversed conditions.*

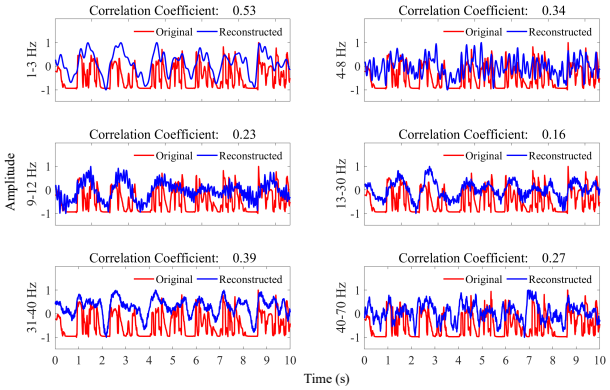

Figure 4: *Examples of original and reconstructed speech envelopes in different frequency bands.*

4. Discussion

4.1. Subject-alignment method works across individuals

Tracking speech envelope based on neural signal is helpful to answer how neural system processes the speech stimuli. In this study, we successfully increased the tracking accuracies of speech envelope using subject-alignment EEG signals (see Figure 2 for the details). The result confirmed our hypothesis that the human brain uses about the same mechanism to process speech stimuli across individuals. The subject-alignment method can decrease the randomly occurred physiological and external noises effectively. In addition, we speculate that it is difficult for subjects to fully focused on the tasks without any distraction during the experiment, which may result in a decrease in task-evoked neural signals. The average of the subject aligned original EEG can reduce the unexpected random actions and increase the certainty of the data. By comparing with single-trial method, the more accurate tracking of speech envelope was obtained using subject-alignment method, which gives a more accurate description for the neural entrainment.

4.2. Neural Entrainment reflects speech intelligibility

Neural entrainment to intelligible speech is stronger than non-intelligible speech in all of the frequency bands, including delta, theta, alpha, beta and gamma bands, as shown in Figure 3. As mentioned previously [5], speech envelope is considered to be related to syllable level of speech. Since the numbers of the syllable boundaries of the original speech and time-reserved speech did not change so much, and the neural entrainment to time-reserved speech is also activated strongly, we can reasonably speculate that the difference in reconstruction accuracies between accessible and inaccessible speech is caused by the sematic processing. That is, the neural entrainment to speech envelope reflects speech intelligibility.

4.3. The proposed method provides an possibility to study gamma band by scalp neural signal

In previous scalp EEG/MEG studies, there is few report about the neural entrainment to speech envelope in gamma bands. Moreover, many studies reported that envelope reconstruction accuracy above 12 Hz is lower than chance level [7, 14, 19]. The main finding of neural response in high frequency gamma band is reported by some intracranial electrography studies, and they indicate that gamma frequency band may represent lexical and linguistic process (see review [20]). However, the intracranial electrography is not friendly to healthy people. Therefore, the finding of neural tracking of speech envelope in gamma band by proposed method provides a possibility to address the neural mechanism of the gamma band by scalp EEG in the future.

5. Conclusions

In this study, we hypothesized human brain function for speech processing is consistent across individuals. Then, we proposed a method to align the subject data for the same stimuli to enhance the neural entrainment to speech in EEG data. Based on the subject-alignment method, the correlation between the reconstructed speech envelope and the original one improved to about 0.5, and the error reduction rate was around 33% comparing with the best envelope reconstruction accuracy of 0.25 in the previous studies [14, 19]. It is found that neural entrainment to speech also occurs in gamma band, which could not be observed in previous studies except for the invasive methods such as intracranial electrography. For the difference between the reconstruction-accuracy of story and time-reversed speech, one reasonable explanation is that neural entrainment involves in the processing of the semantic information. Therefore, the neural entrainment to speech reflects high-level linguistic process to some extent.

However, in this study, we ignore that the time-delay is specific for different subjects. Although, we assume that the TRFs is consistent for all subjects, in fact, the latency of neural response may be not the same across subjects. Therefore, future work should estimate and align the time delays for different subjects respectively, which possibly gets better results.

6. Acknowledgements

This study is supported in part by JSPS KAKENHI Grant (20K11883), and in part by National Natural Science Foundation of China (No.61876126). The authors thank Jinfeng Huang for useful discussions and help on the EEG data collecting.

7. References

[1] G. Zhang, Y. Si, and J. Dang, "Revealing the dynamic brain connectivity from perception of speech sound to semantic processing by eeg," *Neuroscience*, vol. 415, pp. 70–76, 2019.

[2] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.

[3] C. Brodbeck, A. Presacco, and J. Z. Simon, "Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension," *NeuroImage*, vol. 172, pp. 162–174, 2018.

[4] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Current Biology*, vol. 28, no. 5, pp. 803–809, 2018.

[5] X.-Y. Pan, J.-J. Zou, P.-Q. Jin, and N. Ding, "The neural encoding of continuous speech-recent advances in EEG and MEG studies," *Sheng li xue bao:[Acta Physiologica Sinica]*, vol. 71, no. 6, pp. 935–945, 2019.

[6] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.

[7] H. Park, R. A. A. Ince, P. G. Schyns, G. Thut, and J. Gross, "Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners," *Current Biology*, vol. 25, no. 12, pp. 1649–1653, 2015.

[8] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech intelligibility predicted from neural entrainment of the speech envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, no. 2, pp. 181–191, 2018.

[9] M. F. Howard and D. Poeppel, "Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension," *Journal of neurophysiology*, vol. 104, no. 5, pp. 2500–2511, 2010.

[10] R. E. Millman, S. R. Johnson, and G. Prendergast, "The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility," *Journal of cognitive neuroscience*, vol. 27, no. 3, pp. 533–545, 2015.

[11] B. Zoefel and R. VanRullen, "EEG oscillations entrain their phase to high-level features of speech sound," *Neuroimage*, vol. 124, pp. 16–23, 2016.

[12] G. A. o. t. W. M. Association, "World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects." *The Journal of the American College of Dentists*, vol. 81, no. 3, p. 14, 2014.

[13] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[14] O. Etard and T. Reichenbach, "Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise," *The Journal of Neuroscience*, vol. 39, no. 29, p. 5750, 2019.

[15] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Auditory-Inspired End-to-End Speech Emotion Recognition Using 3D Convolutional Recurrent Neural Networks Based on Spectral-Temporal Representation," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[16] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli," *Frontiers in human neuroscience*, vol. 10, p. 604, 2016.

[17] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[18] D. M. Corey, W. P. Dunlap, and M. J. Burke, "Averaging correlations: Expected values and bias in combined Pearson rs and Fisher's z transformations," *The Journal of general psychology*, vol. 125, no. 3, pp. 245–261, 1998.

[19] J. Zou, J. Feng, T. Xu, P. Jin, C. Luo, J. Zhang, X. Pan, F. Chen, J. Zheng, and N. Ding, "Auditory and language contributions to neural encoding of speech features in noisy environments," *NeuroImage*, vol. 192, pp. 66–75, 2019.

[20] A. Kösem and V. Van Wassenhove, "Distinct contributions of low- and high-frequency neural oscillations to speech comprehension," *Language, Cognition and Neuroscience*, vol. 32, no. 5, pp. 536–544, 2017.