The Interplay of Attention and Gaze Direction in EEG and Audio Envelope Analysis

Maryam Bajool, Juan Daniel Galeano Otavaro, Bernhard U. Seeber

Audio-Signalverarbeitung, Technische Universität München, 80333 München, maryam.bajool@tum.de

Introduction

A significant challenge for improving hearing devices remains noisy environments, where multiple individuals engage in simultaneous conversations, the "cocktail party problem." Effectively addressing this challenge involves identifying the specific speaker to whom an individual is listening. Understanding how the brain processes and distinguishes sounds is thus a key to improving hearing aids. Earlier studies have described auditory responses and receptive fields for pure tone to complex stimuli [1, 2]. Researchers have employed various methods, including invasive and non-invasive techniques like electroencephalography (EEG) [3-5]. Despite the inherently nonlinear nature of the human brain, treating the system as a linear model has successfully discerned the target speaker from neural recordings. These methods involve recording brain signals in a two-speaker environment to unveil the brain's ability to track the features of the attended speaker. Encoding speech properties such as envelope, spectrum, and phonetics into neural responses [6, 7], known as the forward method, has been achieved through system identification. Recent research has shown that individuals, when exposed to narratives or sentences, exhibit neural activity that aligns with the speech envelope of the acoustic stimulus [4, 5, 8, 9]. The backward method was introduced such that it would map the multi-channel neural signal to acoustic speech, and it has demonstrated promising results [10]. Using correlation coefficients between predicted and ground truth outputs provides a quantitative and objective means of assessing the accuracy and reliability of both the forward and backward methods. A higher correlation value indicates a strong correspondence between the predicted and actual neural responses in the forward method. It indicates the accurate reconstruction of neural responses from acoustic speech in the backward method.

Moreover, it has been observed that visual input can enhance the neural tracking of acoustic speech [11, 12]. We designed an experiment in anechoic space to study the roles of attention and gaze direction in a two-speaker environment. Employing correlation analysis, the investigation aims to uncover the impact of attention on target envelope features of speech and brain signals using both forward and backward methods. Notably, in addition to attention, we explore the potential impact of gaze direction on the correlation values between predicted and accurate data. By incorporating gaze direction into our correlation analysis, we aim to uncover any distinctive patterns or variations in how individuals direct their attention within the auditory scene.

Materials and Methods

Experimental Procedure

Five native German speakers participated in this preliminary study after giving written informed consent. They were audiometrically confirmed to be of normal hearing. During the experiment, they sat in the center of the Simulated Open Field Environment (SOFE) loudspeaker array in the anechoic chamber at TUM [13] and placed their heads against a headrest to suppress movements. The experiment included two concurrent audio streams from audiobooks, one female narrating the audiobook “Sophie und Hans Scholl,” and one male narrating the audiobook “Ludwig van Beethoven,” both published by Amor Verlag GmbH and used with permission. The audio recordings were available at a 44.1 kHz sampling rate. Silent gaps were truncated to a maximum of 0.5s long, and the data were divided into one-minute-long intervals. Each audio stream was presented at 55 dBA. Audio streams were played from loudspeakers placed at ear height at $-20°$ and $20°$ azimuth, and a 512-taps FIR-filter was used to equalize loudspeakers. The experiment consisted of eight familiarization runs followed by 24 data collection runs. In the SOFE, visual information was displayed with a projector on a screen in front of the subject with a width of 4.3 m and a height of 2.7 m. The subject sat at a distance of 2.4 m from the projection screen and had a visual azimuthal angle coverage of $\pm 45°$. Each trial commenced with the appearance of an arrow for five seconds in the middle of the screen, indicating the direction of the attended speech source. Subsequently, the arrow disappeared, and a fixation gaze point was shown randomly at either the attended or the unattended source position. Subjects were asked to pay attention to the indicated audio stream while fixating on the fixation point during each trial. Simultaneously, EEG data were collected. At the end of each trial, participants answered a four-choice question regarding the attended speech to encourage attention. The directions of target audio, speaker, and gaze fixation placement were randomly chosen for each trial.

Data Collection

The Electroencephalography (EEG) data were recorded from five subjects while listening to the stimuli mentioned above. 64-channel actiCHamp Plus device made by Brain Products (10/20-system) was used to collect data at the rate of 1000 Hz. The channel FCz was selected as a reference.

Data Preprocessing

EEG data were first downsampled to 128 Hz and filtered over the range of $0.5 - 30$ Hz. Data were re-referenced to the average of all channels. The speech envelope was extracted by a Gammatone filter bank [14] with 31 channels spaced between 80 Hz to 8000 Hz by one equivalent rectangular bandwidth. The envelopes were extracted by taking the absolute value and raising it to the power of 0.3 [15]. The speech envelopes were downsampled to 128 Hz. EEG data were analyzed using EEGLAB [16] and Fieldtrip [17] toolboxes.

Stimulus-Response Model (Forward and Backward Models)

Temporal response functions (TRFs) [18] attempt to characterize the neural responses of channel n, denoted as $r(t,n)$, by modeling them as a linear combination of the stimulus, $s(t-\tau)$ at different time delays, τ, with each component weighted by $w(\tau,n)$. The model accounts for any potential noises or errors in the neural responses by $\varepsilon(t,n)$. TRF is also known as the *forward model* or encoder [10]:

$$r(t,n) = \sum_{\tau} w(\tau,n)s(t-\tau) + \varepsilon(t,n). \qquad (1)$$

The model predicts neural responses $\hat{r}(t,n)$ based on minimizing the mean squared error (MSE) between the actual and estimated neural responses:

$$min\ \varepsilon(t,n) = \sum_{t} [r(t,n) - \hat{r}(t,n)]^2, \qquad (2)$$

practically, using reverse correlation with $\mathbf{S}$, which is the lagged time series of s:

$$\boldsymbol{w} = (\mathbf{S}^{\mathbf{T}}\mathbf{S})^{-\mathbf{1}}\mathbf{S}^{\mathbf{T}}\boldsymbol{r}. \qquad (3)$$

To address the challenges of an ill-posed estimation problem and mitigate overfitting due to the autocovariance matrix $\mathbf{S}^{\mathbf{T}}\mathbf{S}$, regularization techniques have been employed. One such method is Ridge regularization, which introduces a hyperparameter known as the "ridge parameter" (λ):

$$\boldsymbol{w} = (\mathbf{S}^{\mathbf{T}}\mathbf{S} + \lambda \boldsymbol{I})^{-\mathbf{1}}\mathbf{S}^{\mathbf{T}}\boldsymbol{r} \qquad (4)$$

This parameter can be fine-tuned during cross-validation to optimize the correlation between the original and predicted responses.

The *backward method* [10, 19] employs a decoder model $g(\tau, \mathrm{n})$ to map the lagged window of neural responses r(t,n) to the audio stimulus s(t). The reconstructed stimulus $\hat{s}$(t) is:

$$\hat{s}(t) = \sum_{n}\sum_{\tau} g(\tau,n)r(t+\tau,n). \qquad (5)$$

The decoder also operates by minimizing the MSE between $s(t)$ and $\hat{s}$(t). The models associate a single stimulus feature to each of the multiple channels. The regularization procedure is similarly applied to the backward method as it is to the forward method, involving the replacement of the autocovariance matrix of $\boldsymbol{R}^T\boldsymbol{R}$ and the audio stimulus $\boldsymbol{s}$. Here, $\boldsymbol{R}$ represents the lagged time series of neural responses.
The implementation of the MATLAB-based mTRF Toolbox [10] was used. The model, in the forward method w and in the backward method g, is obtained to predict the output by training on all data features (e.g., response channels).

Data Analysis

The lag window was set to a length of 500 ms, with $\tau_{min} = -100$ ms and $\tau_{max} = 400$ ms to capture the neural response components of event-related potentials (ERP) [10]. The 24 trials of each subject were divided into two distinct groups: 12 trials when a participant fixated at the same gaze fixation place as the attended audio source and the remaining 12 trials where the fixation was directed towards the opposite direction, denoted as *congruent* and *incongruent* for simplicity. The training procedure employed a "leave-one-out" approach, where a single trial served as the test data, while the rest were utilized for training. The model for each individual training trial was computed, and the average of all models was considered as a final model to predict the unseen test data. Cross-validation was done for the optimization step. Pearson's correlation analysis was employed to assess the performance of these models. In the case of the forward method, the correlation coefficient between the estimated EEG responses and the actual EEG data was calculated for each channel and subsequently averaged across all channels. In the backward method, the predicted audio envelope was correlated with the original audio envelope. Specifically, each trial was evaluated to discern whether the attended correlation value was higher than the unattended correlation value. The resulting binary outcomes were then used to calculate the percentage of correctly classified trials for each subject.

Statistical Analysis

In order to evaluate the variance of obtained correlation values, repeated measures analysis of variance (ANOVA) tests are utilized. Prior to conducting the statistical analysis, Fisher Z-transformation is applied to account for the non-Gaussian distribution of correlation values.

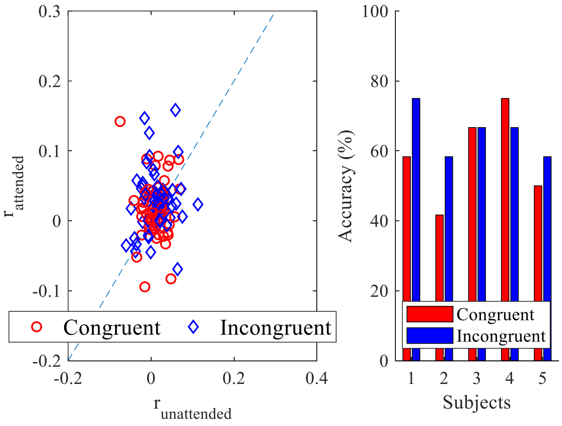

Figure 1: Forward model: (left) Correlation values of the predicted EEG envelope with the observed EEG envelope for each trial, both for the attended and the unattended speech stream, denoted as $r_{attended}$ and $r_{unattended}$. Red points represent correlation values corresponding to the scenario where the fixation point was at the attended loudspeaker, while blue points illustrate values for the case when the gaze point was positioned at the unattended loudspeaker. (right) The corresponding accuracies are presented for each subject.

Results

Figure 1 and Figure 2 display the Pearson correlation values alongside the obtained accuracies of all trials and subjects for forward and backward methods, respectively. Accuracy values express percent correct when selecting the attended speaker based on comparing correlation values of attended and unattended. The y-axis represents the calculated correlation values for attended envelope decoding ($r_{attended}$), while the x-axis corresponds to the values of unattended envelope decoding ($r_{unattended}$). Correlation values for *congruent* are shown in red, and the results of *incongruent* are shown in blue. Our preliminary data of five subjects demonstrate the impact of attention on decoding accuracy, as evidenced by the predominance of higher $r_{attended}$ compared to $r_{unattended}$. Figures 1 and 2 reveal that the distribution of correlation points is primarily situated above the identity line, indicating closer proximity to the attended values. This effect is especially evident in the **backward** method, as illustrated in Figure 2. Upon comparing the value points in red and blue, no significant differences were identified in the distribution of correlation values for each case, indicating that gaze direction had no discernible effect (Forward method: $p = 0.23477$, Backward method: $p = 0.43279$).

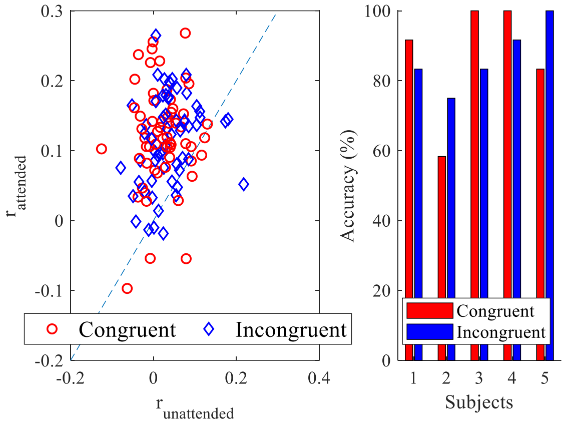

Figure 2: Backward model: The explanation as Figure 1, but for the backward model.

Conclusion

In this study, we examined the roles of attention and gaze direction. Preliminary findings derived from the study comprising five participants suggest a noticeable impact of attention on EEG signals, which can be readily recovered with linear methods. Gaze direction did not affect EEG decoding accuracy. This suggests that attention can be measured without the influence of gaze direction in multi-talker situations.

References

[1] B. M. Calhoun and C. E. Schreiner, "Spectral envelope coding in cat primary auditory cortex: linear and non-linear effects of stimulus characteristics," (in eng), no. 0953-816X (Print).

[2] S. A. Shamma, H. Versnel, and N. Kowalski, "Ripple analysis in ferret primary auditory cortex. I. Response characteristics of single units to sinusoidally rippled spectra," 1994.

[3] J. A. O'sullivan *et al.*, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral cortex,* vol. 25, no. 7, pp. 1697-1706, 2015.

[4] L. Decruy, J. Vanthornhout, and T. Francart, "Hearing impairment is associated with enhanced neural tracking of the speech envelope," *Hearing Research,* vol. 393, p. 107961, 2020/08/01/ 2020, doi: 10.1016/j.heares.2020.107961.

[5] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "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. 25, no. 5, pp. 402-412, 2017, doi: 10.1109/TNSRE.2016.2571900.

[6] W. Biesmans, J. Vanthornhout, J. Wouters, M. Moonen, T. Francart, and A. Bertrand, "Comparison of speech envelope extraction methods for EEG-based auditory attention detection in a cocktail party scenario," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* 25-29 Aug. 2015 2015, pp. 5155-5158, doi: 10.1109/EMBC.2015.7319552.

[7] B. Khalighinejad, G. Cruzatto da Silva, and N. Mesgarani, "Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech," (in eng), *J Neurosci,* vol. 37, no. 8, pp. 2176-2185, Feb 22 2017, doi: 10.1523/jneurosci.2383-16.2017.

[8] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: functional roles and interpretations," (in English), *Frontiers in Human Neuroscience,* Review vol. 8, 2014-May-28 2014, doi: 10.3389/fnhum.2014.00311.

[9] E. C. Lalor, A. J. Power, R. B. Reilly, and J. J. Foxe, "Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli," *Journal of Neurophysiology,* vol. 102, no. 1, pp. 349-359, 2009, doi: 10.1152/jn.90896.2008.

[10] 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.

[11] M. J. Crosse, G. M. Di Liberto, and E. C. Lalor, "Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration," *Journal of Neuroscience,* vol. 36, no. 38, pp. 9888-9895, 2016.

[12] F. Ahmed, A. R. Nidiffer, A. E. O'Sullivan, N. J. Zuk, and E. C. Lalor, "The integration of continuous audio and visual speech in a cocktail-party environment depends on attention," *NeuroImage,* vol. 274, p. 120143, 2023.

[13] B. U. Seeber and S. W. Clapp, "Interactive simulation and free-field auralization of acoustic space with the rtSOFE," *The Journal of the Acoustical Society of America,* vol. 141, no. 5, pp. 3974-3974, 2017.

[14] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE,* 1987, vol. 2, no. 7.

[15] D. D. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. De Cheveigne, "A comparison of regularization methods in forward and backward models for auditory attention decoding," *Frontiers in neuroscience,* vol. 12, p. 531, 2018.

[16] 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.

[17] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Computational intelligence and neuroscience,* vol. 2011, pp. 1-9, 2011.

[18] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic

listening," (in eng), *J Neurophysiol*, vol. 107, no. 1, pp. 78-89, Jan 2012, doi: 10.1152/jn.00297.2011.

[19] S. Haufe *et al.*, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96-110, 2014/02/15/ 2014, doi:10.1016/j.neuroimage.2013.10.067.

[20] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of neurophysiology*, vol. 85, no. 3, pp. 1220-1234, 2001.