Interspeech 2024
1-5 September 2024, Kos, Greece

Auditory Attention Decoding in Four-Talker Environment with EEG

Yujie Yan[2,3], *Xiran Xu*[1,3], *Haolin Zhu*[1,3], *Pei Tian*[1], *Zhongshu Ge*[1], *Xihong Wu*[1,3], *Jing Chen*[1,2,3]

[1]Speech and Hearing Research Center, School of Intelligence Science and Technology, Peking University, China [2]National Biomedical Imaging Center, College of Future Technology, Peking University, China [3]National Key Laboratory of General Artificial Intelligence, China

janechenjing@pku.edu.cn

Abstract

Auditory Attention Decoding (AAD) is a technique that determines the focus of a listener's attention in complex auditory scenes according to cortical neural responses. Existing research largely examines two-talker scenarios, insufficient for real-world complexity. This study introduced a new AAD database for a four-talker scenario with speeches from four distinct talkers simultaneously presented and spatially separated, and listeners' EEG was recorded. Temporal response functions (TRFs) analysis showed that attended speech TRFs are stronger than each unattended speech. AAD methods based on stimulus-reconstruction (SR) and cortical spatial lateralization were employed and compared. Results indicated decoding accuracy of 77.5% in 60s (chance level of 25%) using SR. Using auditory spatial attention detection (ASAD) methods also indicated high accuracy (94.7% with DenseNet-3D in 1s), demonstrating ASAD methods' generalization performance.

Index Terms: auditory attention decoding, auditory spatial attention detection, stimulus reconstruction, temporal response functions, EEG, DenseNet

1. Introduction

In complex auditory scenes with multi-talker speaking simultaneously, a listener with normal hearing can focus on a target speaker while ignoring others, which is known as the "cocktail party problem" [1]. Studies in neuroscience show that auditory selective attention significantly enhances the cortical speech-envelope tracking of the attended stream, and hence the attended speech could be identified with cortical neural responses [2] by the technology of auditory attention decoding (AAD). This technology has the potential to be used in developing neurologically controlled hearing devices [3, 4].

Most studies of AAD methods achieved a high decoding accuracy in two-talker scenarios [2, 5, 6], in which subjects were instructed to attend one of the two presented talkers while ignoring the other. However, there is a significant gap between the two-talker scenarios and real-world settings. Few studies discussed how well these AAD methods results generalize to multi-talker acoustic scenarios [7]. Schäfer PJ et al. [8] collected an EEG database for AAD in an environment having four spatially separated talkers, and they employed the stimulus-reconstruction method (SR, [2]) on this database. Their results showed that SR correctly classified the attended talker on an average accuracy of 61.1% of the trials in 120s, well above the chance level of 25%. However, it remains unclear: 1) what the characteristics of cortical neural responses and, 2) how to improve the AAD accuracy, in such four-talker environment.

Although SR has been widely used in AAD, the auditory spatial attention detection (ASAD) methods outperformed significantly in environment where speech streams were localized from different spatial places, which capitalized on the neural encoding of auditory spatial attention through brain lateralization [9]. Existing ASAD algorithms can be categorized into traditional decoders with a feature extraction frontend and pattern classification backend, i.e., the common spatial pattern (CSP, [10]), the Riemannian geometry-based classifier (RGC, [11]), and DNN-based decoders, such as the STAnet [12], LSM [13], XAnet [14], DenseNet [15]. Generally speaking, the DNN-based decoders outperformed the others. However, all those DNN-based ASAD models were evaluated on the database which was collected in a two-talker scenario, the performance in four-talker scenarios is still unclear.

To investigate the performance of SR-AAD and DNN-based ASAD in multi-talker scenarios, a spatially separated four-talker environment was set up and the EEG was recorded when subjects attended to one of the four talkers. The temporal response functions (TRFs, [16, 17]) were used to analyze the dynamic properties of cortical envelope tracking activities in this four-talker environment. The difference in TRFs between the attended and the unattended speech was compared. SR was adopted to obtain a direct measure of the cortical tracking to the attended speech and was used as a baseline decoding method.

Due to the superior ability of deep convolutional neural networks to extract spatial and temporal features of EEG [15, 18, 19], four related models [15] (CNN-baseline, CNN-3D, DenseNet-3D, DenseNet-3D with bootstrapping) were used for the ASAD task with the new four-talker database. The implementation code and database are available on Github: `https://github.com/xuxiran/AAD_4direction_code.git` and Zenodo: `https://zenodo.org/records/10803229`.

2. Materials and Methods

2.1. Participants

Sixteen university students(age range: 19–26 years) with normal hearing took part in the experiment, whose whose audiometric thresholds were less than 20 dB hearing level at frequencies from 250 Hz to 8000 Hz for both ears. All of the subjects were Mandarin-native speakers and right-handed, and ten of them were male. Before the experiments, the subjects were informed about the procedure and the objectives of the experiment. The design of the experiment was planned in accordance with ethics guidelines and was approved by the Peking University Institutional Review Board.

10.21437/Interspeech.2024-739

2.2. Stimuli and experimental procedure

The audio materials used in the experiment were proposed in the previous study [5]. The speech corpus was selected from the book (Chinese translation) *Twenty Thousand Leagues under the Sea* by Jules Verne. Two speakers (one female and one male, standard-Mandarin speakers) narrated twenty different segments, and each lasted for more than 60s. The mean F0 was about 207 Hz and 124 Hz for the two speakers. The other two speakers' speeches were produced by shifting up the F0 of the original speech with the speech synthesis technology of Adobe Audition software, resulting in a mean F0 of 230 Hz and 136 Hz for the female and the male speech, respectively. All speeches were clipped into consecutive 60s segments starting from the speech onset. In the experiment, 40 speech combinations were used, with each consisting of speech segment of one female, one male, one female with a higher F0 and one male with a higher F0. All the audio was recorded at a sampling rate of 48 kHz.

The schematic of the acoustic environment used for this study was similar to previous work [8], as shown in Figure 1. The experiments were conducted on an anechoic chamber, dimensions of which were 6.5 m×4.8 m×3.2 m [20], with a cutoff frequency of 70 Hz. A circular acoustic free-field system, consisting of four loudspeakers (Dynaudio BM 6A) with a radius of 1.6m, was established inside the room. The loudspeakers were symmetrically positioned in a semicircle at equal angular distances, with locations at +30° (LS1), -30° (LS2), +90° (LS3), and -90° (LS4), aligned with the height of the listener's ears.

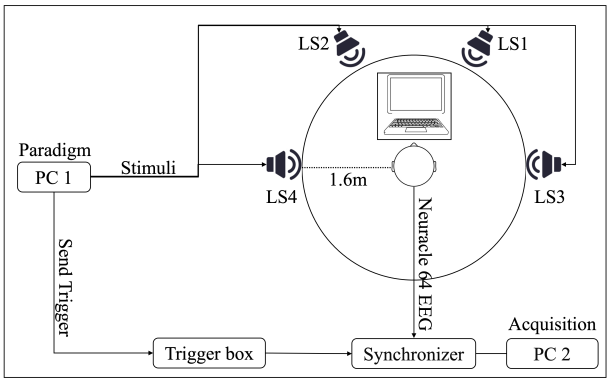

Figure 1: *Setup of the stimulus presentation and the data acquisition.*

In the experiment, subjects sat right in front of a desk with their head fixed by a chinrest in the center of the circle. A monitor was on the desk to present subjects' visual feedback and relevant instructions. Before the stimulus presentation, subjects were cued to attended to the speech presented by a loudspeaker with a certain direction and ignore the other three. During the presentation, the speeches of the four talkers in a speech combination were presented through the four loudspeakers at 55 dBLA_{eq}, respectively. For each subject, 10 combinations out of the 40 were assigned to each space condition, and each combination was presented once with different attended talker. Thus, each subject undertook 40 trials in total.

2.3. EEG data acquisition and preprocessing

The EEG signal was collected by a 64-channel EEG amplifier (NeuSen Wireless EEG/ERP, Neuracle, China), and the recording software (NeuSen Recorder, Neuracle, China) recorded the EEG signal at a sampling rate of 1000 Hz. The placement of the 64 electrodes was facilitated by an EEG recording cap (Neuracle, China), adhering to the international 10-20 system for electrode positioning.

Due to the two different decoding methods, two processing procedures were adopted. For SR, following the previous studies [5, 21], the EEG data were re-referenced, down-sampled to 64Hz and baseline corrected. Since the low-frequency (< 8 Hz) neural activity in the auditory cortex was reported phase-locked to speech envelopes [2], the EEG data were subsequently bandpass filtered between 2 and 8 Hz. For ASAD, the EEG data was down-sampled to 128 Hz, bandpass filtered between 14 and 31 Hz, and normalized, as the previous studies did [15]. EEG data preprocessing was performed using the EEGLAB toolbox [22] on MATLAB.

2.4. Speech temporal envelope extraction

The procedure for speech temporal envelope extraction was the same as previous works [5, 21]. Each clean speech waveform was processed through a filter bank consisting of 8 bandpass filters, simulating the bandpass characteristics of the basilar membrane. The center frequencies of these filters spanned from 150Hz to 8000Hz, evenly spaced on the equivalent rectangular bandwidth (ERB) scale. Then, a Hilbert transform was applied to the output of each filter to obtain the resultant analytical envelope. To account for the cochlear compression nonlinearity, these signals were raised to a power of 0.3 [23], and subsequently low-pass filtered at a cutoff frequency of 8Hz. Finally, the envelopes of all frequency bands were averaged.

2.5. TRFs estimation

The TRFs estimation methods were the same as previous work [21]. TRFs analysis was conducted using the mTRF toolbox [24]. For speech temporal envelope $s(t)$ sampled at discrete time $t(t = 1, \cdots, T)$ and the corresponding EEG $r(t, n)$ recorded at channel $n(n = 1, \cdots, N)$, suppose that a set of spatio-temporal filters (i.e., the channel-specific TRFs) $w(\tau, n)$ could map from $s(t)$ to $r(t, n)$ in a convolutional way, as in equation 1:

$$\hat{r}(t, n) = \sum_{\tau} \mathrm{w}(\tau, \mathrm{n})\mathrm{s}(t - \tau) \tag{1}$$

where $\hat{r}(t, n)$ represents estimated EEG. To calculate the TRFs for the attended and the unattended speech, a leave-one-out cross-validation approach was utilized to fine-tune the ridge parameter λ. This involved predicting EEG data for each trial using the averaged TRFs from all other trials.

2.6. Speech stimulus reconstruction

The SR procedures were the same as previous work [5, 8]. Similar to TRFs, a set of filters were used to map EEG to speech envelope. Suppose that a set of spatio-temporal filters $g(\tau, n)$ represent the linear backward mapping from $r(t, n)$ to $s(t)$, was shown in equation 2:

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

where $\hat{s}(t)$ represents the reconstructed speech envelope. The spatio-temporal filters (also called decoders) integrate the neural responses across a certain range of time lag (τ = 500ms

post), then the integrated signals are summed across channels to obtain the reconstructed speech envelope. Similar with TRFs estimation, the decoders were also trained using ridge regresson with a coross-validation approach. More details were referred to [5]. For each trial, the reconstructed envelope was calculated through the decoder, and correlation coefficient (Pearson's r) between reconstructed speech envelope and the attended speech envelope was calculated. An attended speech stream was considered as correctly-identified when the correlation between the reconstructed envelope and the actual attended speech envelope was greater than that between the reconstructed envelope and the other unattended speech envelopes. The AAD accuracy was evaluated by the percentage of correctly-identified trials for each subject.

2.7. 3D Representation of EEG

The present study used 3D representation in ASAD which has shown effective in representing spatial features. The definition of 3D Representation of EEG was similar to previous study [23]. Let $E \in \mathbb{R}^{C\times T}$ be the EEG signals for a decision window, with C channels and T time samples. The network takes E as the input and makes auditory spatial attention decisions. To get the 3D representation of EEG, the channel indexes are converted from 1D (with the number of C) to 2D ($H \times W$) in the same manner used in [15]. The EEG signals are transformed from $E \in \mathbb{R}^{C\times T}$ into $E \in \mathbb{R}^{H\times W\times T}$ using the topological location of EEG channels in space and the blank positions are filled with zeros [15], where H and W represent the spatial (channel) dimension, and T represents the temporal dimension.

2.8. ASAD models

Four DNN-based models [15] were implemented to validate the feasibility of the ASAD method on this database.

The first one was a CNN model (CNN-baseline) [25]. A single 2D convolutional layer (kernel size: 64 $\times$ 17, channel number: 5) was used to extract features from the 2D representation of EEG data $E \in \mathbb{R}^{C\times T}$. ReLU activation function was then used followed by average pooling over the temporal dimension. Finally, a classification head with two fully connected layers were used to output the spatial attention.

The second model was a simple CNN-3D model [15]. The input of this model was the 3D representation of EEG data $E \in \mathbb{R}^{H\times W\times T}$, where H = 9, W = 9 and T = 128 when the decision window was 1 second. In the CNN-3D model, twenty independent 5 $\times$ 5 spatial filters were used to extract features. ReLU was used as the activation function after the convolution step followed by average pooling over the temporal dimension and the two fully connected layers. The CNN-3D model was very similar to the CNN-baseline except it provided EEG data with spatial distribution of EEG electrodes and used simple spatial filters to extract features.

The third and last models utilized DenseNet-3D [15]. At specific sampling time of 3D representation of EEG$\in \mathbb{R}^{H\times W\times T}$, corresponded to $E_t \in \mathbb{R}^{H\times W}$ which is a 2D topography and as the input of the DenseNet-2D. The DenseNet-2D architecture consisted of a convolutional layer, a max-pooling layer, dense blocks, transition layers, global average pooling, a fully connected layer, and a softmax classifier. The output of the classifier was a four-class decision, and the values, 1, 2, 3, and 4, indicated the attended location with the EEG input at time t, E_t, decoded as +30°, -30°, +90°, -90°, respectively. To extract temporal and spatial features of EEG signals, the DenseNet-2D was transferred to DenseNet-3D by inflating its filters and pooling kernels, more details were referred to [15]. The difference between the third and last model was that the former was without bootstrapping signifying the absence of initializing parameters from DenseNet-2D.

2.9. Implementation details for ASAD

The evaluation of the ASAD models was carried out with a 5-fold cross-validation paradigm. Specifically, the EEG data of each spatial attention condition were divided into five groups at the trial level, with each group consisting of 2 trials from each condition. Four groups among the five were chosen as the training data and the remaining one was used as the test set. The training data were then split into training and validation subsets in a ratio of 4:1. The trials EEG data in each set were further split into 1s windows without overlapping. The ASAD accuracy was calculated as the percentage of windows with correct decisions out of all windows.

A DenseNet-2D model was trained first, and the model with best accuracy on validation set was used to initialize the weights of the 3D model by bootstrapping [26]. All models were trained and evaluated for individual subjects.

The models were implemented using Pytorch and trained on an NVIDIA RTX 3090 GPU. The cross-entropy loss function was used and an Adaptive Moment Estimation (Adam) optimizer [27] with a learning rate of 10^{-3} was adopted.

3. Results and discussion

3.1. TRFs estimation

To examine the effect of the four-talker environment with spatial separation on cortical processing of speech, average TRFs for both attended and unattended speech were estimated across all subjects and trials. Figure 2 displays the average estimated TRFs for both attended and unattended speech at electrodes covering frontocentral (FC5, FC6) scalp regions. It shows the results of averaging TRFs calculated separately for each of the three unattended speeches, with the TRFs of the remaining three unattended speeches also plotted in the figure as colored dashed lines. It was observed that the TRFs response to attended speech was higher than that to unattended TRFs, similar to previous findings in two-talker scenarios [6, 21]. The finding suggests that, in four-talker scenarios, the neural tracking of the attended speech's envelope was more pronounced than the tracking of each unattended speech.

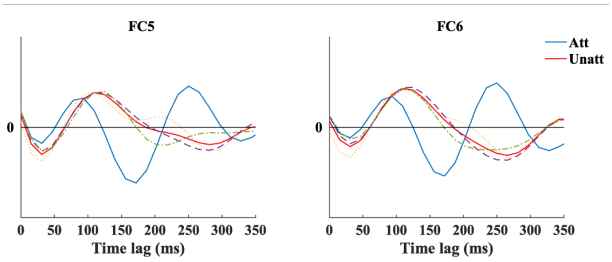

Figure 2: *The TRFs at FC5 and FC6 electrodes for the attended and the unattended speech. TRFs averaged over the other three unattended TRFs. TRFs of the other three unattended speeches are plotted as colored dashed lines.*

3.2. Speech stimulus reconstruction

The speech temporal envelope was reconstructed from the EEG data for each trial, and the correlation coefficient (Pearson's r) between the reconstructed envelope and the envelope of any speech stream was calculated to evaluate the cortical envelope tracking to the attended and the unattended speech. The averaged correlation coefficient between the reconstructed envelope and the envelope of the attended stream for all subjects was 0.0550 (SD: 0.0345). The result of the correlation coefficient was similar to previous studies [8].

Consistently, the reconstructed envelope was more correlated with the attended speech than the other unattended speech. Figure 3 shows the average decoding accuracy for each decision window, indicating that the effectiveness by using SR.

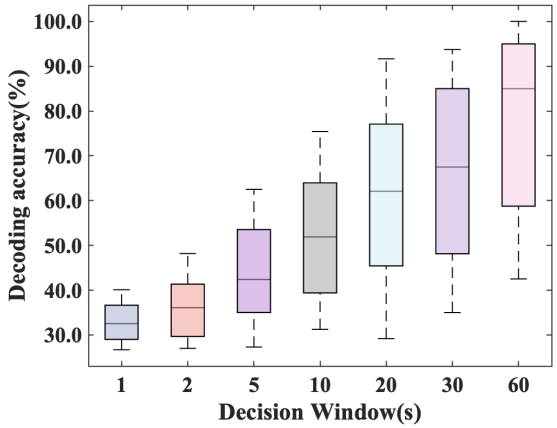

Figure 3: *Decoding accuracy with different decision window by using SR.(chance level of 25%)*

3.3. ASAD models

The ASAD accuracy of the four models across all subjects on this database is shown in Figure 4. The average ASAD accuracy with the 1-second decision window for the four models were all higher than the chance level. The CNN-baseline model reached an accuracy of 61.9±11.9%, lower than that of the CNN-3D model ($t = 10.91, p < 10^{-7}$), indicating the benefit of EEG 3D representation. The DenseNet-3D without bootstrapping achieved an accuracy of 88.7±11.0%, significantly higher than that of the CNN-3D model ($t = 2.58, p = 0.02$), suggesting the advantages of deep network and dense connection. With bootstrapping, the performance DenseNet-3D model was further improved ($t = 2.32, p = 0.035$), of which an accuracy of 94.7± 2.9% was reached. This result verified the effectiveness of bootstrapping. Those results were consistent to the results in the two-talker scenario (84.8±10.4%, 88.6±10.0%, 91.8±7.1%, 94.3±5.7% [15] in KUL database [28]). The results further demonstrated the ability of ASAD methods to detect auditory spatial attention in a more complex environment and the superior performance of DenseNet-3D in complex scenarios and its generalization ability.

3.4. Limitations

Despite of high decoding accuracy, the 64-channel scalp EEG was used in this study, making it unsuitable for further application. Therefore, AAD with fewer electrodes of EEG needs to

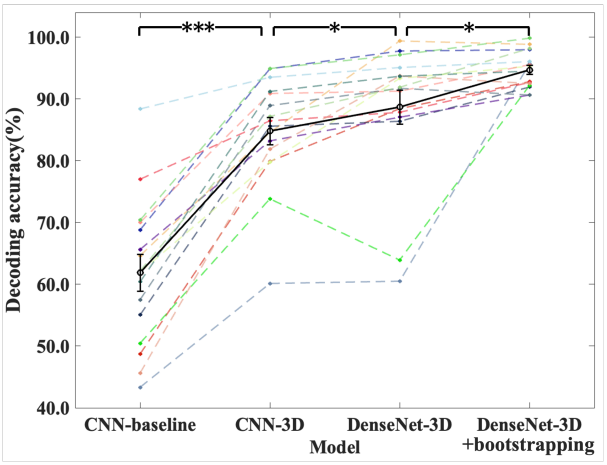

Figure 4: *Decoding accuracy for each of the four models among all subjects with a 1-second decision window. The dashed colorful lines represent the result of individual subjects, and the bold black line represents the average accuracy among subjects.* $*p < 0.05, **p < 0.001$.

be furtherly investigated,so the technology can be incorporated with portable devices, i.e. hearing aids. Besides, the experiment was conducted in an anechoic chamber without reverberation, and only a four-talker setting was designed in this work, but other realistic cocktail party scenarios full of talker uncertainty and room reverberation need to be considered. Given the previous finding [7] that different reverberation environments had varying effects on neural representation, the generalization to real scenarios needs to be carefully treated. Finally, considering that the high decoding accuracy method DenseNet-3D is a deep convolutional neural network, the computational cost is likely much higher than that of baseline models. Model distillation could be employed to reduce the model parameters for future deployment[15].

4. Conclusions

In this work, a new AAD database of a four-talker scenario is introduced, in which the four talkers are spatially separated. The listener's EEG was recorded with the typical auditory attention paradigm. The result of TRFs and SR suggest the attended speech TRFs are stronger than each unattended speech, and the decoding accuracy is 77.5% in 60s which was significantly higher than the chance level. Moreover, with the DNN-based ASAD methods, the decoding accuracy reached 94.7% in the 1s decoding window.

5. Acknowledgement

This work is supported by the National Key Research and Development Program of China (No.2021ZD0201503), a National Natural Science Foundation of China (No.12074012), the High-performance Computing Platform of Peking University and Neuracle Technology (Changzhou) Co., Ltd.

6. References

[1] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, Sep. 1953.

[2] 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, pp. 1697–1706, Jul. 2015.

[3] E. Ceolini, J. Hjortkjær, D. D. E. Wong, J. O'Sullivan, V. S. Raghavan, J. Herrero, A. D. Mehta, S.-C. Liu, and N. Mesgarani, "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, p. 117282, Dec. 2020.

[4] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-Informed Attended Speaker Extraction From Recorded Speech Mixtures With Application in Neuro-Steered Hearing Prostheses," *IEEE Transactions on Biomedical Engineering*, vol. 64, pp. 1045–1056, May 2017.

[5] Z. Fu, X. Wu, and J. Chen, "Congruent audiovisual speech enhances auditory attention decoding with EEG," *Journal of Neural Engineering*, vol. 16, p. 066033, Nov. 2019.

[6] B. Wang, X. Xu, Y. Niu, C. Wu, X. Wu, and J. Chen, "EEG-based auditory attention decoding with audiovisual speech for hearing-impaired listeners," *Cerebral Cortex*, vol. 33, pp. 10 972–10 983, Nov. 2023.

[7] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435–444, Aug. 2017.

[8] P. J. Schäfer, F. I. Corona-Strauss, R. Hannemann, S. A. Hillyard, and D. J. Strauss, "Testing the Limits of the Stimulus Reconstruction Approach: Auditory Attention Decoding in a Four-Speaker Free Field Environment," *Trends in Hearing*, vol. 22, p. 233121651881660, Jan. 2018.

[9] A. Bednar and E. C. Lalor, "Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG," *NeuroImage*, vol. 205, p. 116283, Jan. 2020.

[10] S. Geirnaert, T. Francart, and A. Bertrand, "Fast EEG-Based Decoding Of The Directional Focus Of Auditory Attention Using Common Spatial Patterns," *IEEE Transactions on Biomedical Engineering*, vol. 68, pp. 1557–1568, May 2021.

[11] S. Geirnaert, T. Francart, and A. Bertrand, "Riemannian Geometry-Based Decoding of the Directional Focus of Auditory Attention Using EEG," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 1115–1119.

[12] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, "STAnet: A Spatiotemporal Attention Network for Decoding Auditory Spatial Attention From EEG," *IEEE Transactions on Biomedical Engineering*, vol. 69, pp. 2233–2242, Jul. 2022.

[13] Y. Zhang, H. Ruan, Z. Yuan, H. Du, X. Gao, and J. Lu, "A Learnable Spatial Mapping for Decoding the Directional Focus of Auditory Attention Using EEG," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[14] S. Pahuja, S. Cai, T. Schultz, and H. Li, "XAnet: Cross-Attention Between EEG of Left and Right Brain for Auditory Attention Decoding," in *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, Baltimore, MD, USA, Apr. 2023, pp. 1–4.

[15] X. Xu, B. Wang, Y. Yan, X. Wu, and J. Chen, "A DenseNet-Based Method for Decoding Auditory Spatial Attention with EEG," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul,Korea, Apr. 2024, pp. 1946–1950.

[16] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 11 854–11 859, Jul. 2012.

[17] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European Journal of Neuroscience*, vol. 31, pp. 189–193, Jan. 2010.

[18] Z. Qiu, J. Gu, D. Yao, and J. Li, "Exploring Auditory Attention Decoding using Speaker Features," in *Interspeech 2023*, Dublin, Ireland, Aug. 2023, pp. 5172–5176.

[19] X. Xu, B. Wang, Y. Yan, H. Zhu, Z. Zhang, X. Wu, and J. Chen, "Convconcatnet: a deep convolutional neural network to reconstruct mel spectrogram from the eeg," *arXiv preprint arXiv:2401.04965*, Jan. 2024.

[20] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance-Dependent Head-Related Transfer Functions Measured With High Spatial Resolution Using a Spark Gap," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1124–1132, Aug. 2009.

[21] Z. Fu and J. Chen, "Congruent audiovisual speech enhances cortical envelope tracking during auditory selective attention," in *Interspeech 2020*, Shanghai, China, Oct. 2020, pp. 116–120.

[22] 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, pp. 9–21, Mar. 2004.

[23] C. J. Plack, A. J. Oxenham, A. M. Simonson, C. G. O'Hanlon, V. Drga, and D. Arifianto, "Estimates of compression at low and high frequencies using masking additivity in normal and impaired ears," *The Journal of the Acoustical Society of America*, vol. 123, pp. 4321–4330, Jun. 2008.

[24] 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, Nov. 2016.

[25] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *eLife*, vol. 10, p. e56481, Apr. 2021.

[26] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Jul. 2017, pp. 4724–4733.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, Dec. 2014.

[28] 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, pp. 402–412, May 2017.