INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Exploring Auditory Attention Decoding using Speaker Features

Zelin Qiu[1,2], *Jianjun Gu*[1], *Dingding Yao*[1], *Junfeng Li*[1,*]

[1]Institute of Acoustics, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China
{qiuzelin, gujianjun, yaodingding}@hccl.ioa.ac.cn, junfeng.li.1979@gmail.com

Abstract

The auditory attention decoding (AAD) approach aims to determine the identity of the attended talker in a multi-talker scenario using neuro recordings. In the past few years, various AAD methods have been proposed, and most of them rely on speech envelope reconstruction, which unfortunately face challenges with shortened decoding windows. Inspired by the findings that voices with different acoustic features arouse diverse brain activities in a very short period, this paper proposes to use speaker voice features instead of speech envelope as a speaker indicator for conducting AAD in short-time situations. To achieve this, a novel dual-branch convolutional network (DBCNet) is proposed to estimate speaker features from EEG. Results show that the proposed method achieves higher decoding accuracy than existing methods for short decoding windows (approximately 75% for 0.3-s window and 82% for 1.0-s window).

Index Terms: EEG, auditory attention decoding, speaker feature, dual branch convolutional network

1. Introduction

Human are able to follow a specific speaker of interest amidst interference sources. Such phonemenon is called "cocktail party effect" [1] and it could be attributed to the auditory attention during speech perception and the cognitive control in the human brain [2, 3]. Auditory attention is a cognitive process that involves directing cortical processing resources to the most relevant sensory information, and several studies have demonstrated that the attention-driven response to a target speech can be decoded from the cortical responses in the human brain [4, 5]. Due to the enormous potential of neuro-steered hearing aids, the technology of auditory attention decoding (AAD), which uses non-invasive brain recordings to identify which speaker a listener is attending to, has recently come into focus [5, 6, 7, 8].

The generally used AAD methods can be divided into two categories: classification-based and regression-based methods. [9]. In classification-based approaches, the attention is directly predicted in an end-to-end fashion. However, the underlying characteristics of the classification basis cannot be clearly explained, and their generalization ability remains to be confirmed [10]. Regression-based methods can be further divided into forward and backward models, depending on the mapping direction between EEG and auditory stimulus. Specifically, forward models estimate the EEG responses from auditory stimulus, in contrast, backwad models, also known as stimulus reconstruction approaches, resonstruct speech envelopes from EEG signals [11]. The target speaker is then determined according to the correlation between estimated and actual envelopes of the competing speech. For AAD, backward decoding models have been demonstrated to outperform forward models [9], therefore, current research focus has shifted to backward models [11]. Although significant progress has been made in recent years on stimulus reconstruction methods [5, 6, 12, 13, 14], their decoding accuracy is still limited by the properties of speech envelopes. Specifically, as the duration of speech stream becomes shorter, the distinction between the envelopes of the speech and its competing speech becomes more difficult [8], necessitating the use of longer time windows for precise decoding. This is impractical for real-time applications [14]. Consequently, it is necessary to explore other speaker indicators beyond speech envelopes to conduct AAD.

While speech envelopes primarily carry information on speech contents [15], it has been shown that speaker identity analysis occurs in the brain when people attend to a speaker's speech [16]. Furthermore, studies on the hierarchy and time course of voice signal perception [17, 18, 19, 20] suggest that voice feature based identity analysis is processed at a lower processing level and more quickly than speech content analysis.

In light of these findings, we propose to employ speaker voice features instead of speech envelopes to indicate the speaker in the AAD task. To accurately extract speaker features from EEG, we futher propose a novel dual branch convolutional network (DBCNet) that accepts EEG data in both the time and frequency domains as input. We posit that incorporating EEG spectra is beneficial since some noise could be addressed more conveniently in the frequency domain. By employing speaker features and DBCNet, the proposed AAD approach offers two advantages over existing methods. First off, it can handle short-duration speech streams more effectively because it does not rely on the accurate reconstruction of speech envelopes. Second, the DBCNet can leverage the frequency-domain information in the EEG data to improve the accuracy of speaker feature estimation. Overall, experimental results demonstrate that the proposed method achieves higher decoding accuracy than the models based on stimulus reconstruction, which represents a promising approach to AAD that addresses the limitations of existing methods.

2. Problem Formulation

We consider the 2-talker case in this study. Let $\mathbf{s}_a$ and $\mathbf{s}_u$ denote the attended and unattended speech stream observed by the listener, respectively. The corresponding EEG recordings of the listener can be modeled as:

$$E = \mathfrak{A}(\mathbf{s}_a, \mathbf{s}_u) + \mathbf{n} \tag{1}$$

[*] Correspoonding author

10.21437/Interspeech.2023-414

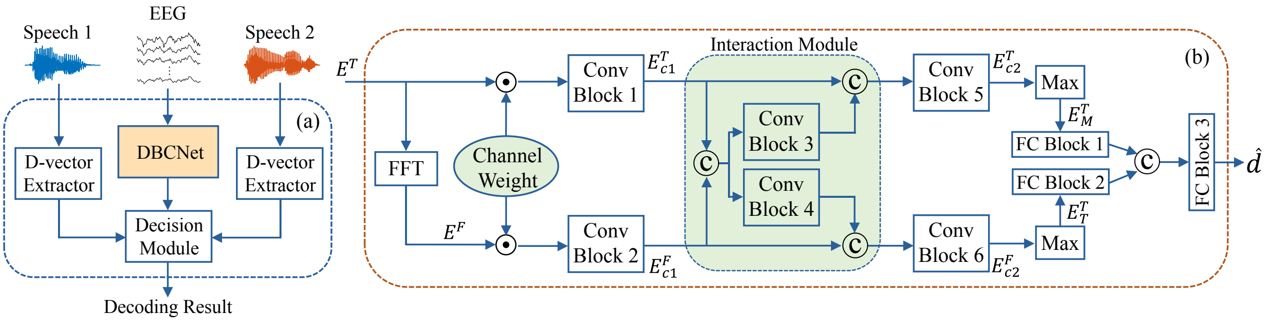

Figure 1: *(a) An overview of the proposed arhitecture. The d-vector extraction module takes speech signals as input and outputs d-vectors correspondingly. The DBCNet maps EEG signals to the d-vector of the target speaker. (b) Proposed DBCNet. The channel number of the three convolutional layers in the each branch is {4, 4, 256}. The kernel size and padding are respectively set as {(5, 5), (5, 65), (32, 65)} and {(2, 2), (2, 0), (0, 0)}. The strides in all convolutional layer are set as (1, 1). The input and output size of the two fully connected layers are set as {(256, 256), (2, 1)}.*

where $\mathfrak{A}(\cdot)$ refers to the auditory pathway and $\mathbf{n}$ are the auditory irrelevant EEG signals. The goal of the AAD task is to develop a decoder $\mathfrak{D}(\cdot)$ to reconstruct the speech features of the attended source, $\mathbf{s}_a$, from the EEG signals, so that:

$$\hat{\mathbf{d}} = \mathfrak{D}(E) \tag{2}$$

To accurately determine the speaker identity, the reconstructed features should satisfy:

$$\mathrm{dis}(\hat{\mathbf{d}}, \mathbf{d}_{s_a}) < \mathrm{dis}(\hat{\mathbf{d}}, \mathbf{d}_{s_u}) \tag{3}$$

where $\mathbf{d}$ is the feature from speech $\mathbf{s}$ such as envelope, spectrum, specker feature, etc. $\mathrm{dis}(\cdot)$ refers to the measurement of distance such Euclidean distance.

3. Model Description

we propose a novel AAD architecture, as illustrated in Fig. 1 (a), which inventively incorporates speaker features into AAD. Our design is made up of three main parts: a feature extractor to obtain speaker features from speech, a DBCNet to estimate speaker features from EEG, and a decision module to choose the target speaker. The speaker feature we use is the d-vector [21], which can be extracted using a rather simple network and has been demonstrated to perform well in speaker verification tasks. Given that we utilize the same d-vector extractor as in [21], we don't go into more detail on this component in this study.

3.1. Dual Branch Convolutional Network (DBCNet)

Due to the ability to extract spatial features, convolutional neural networks (CNNs) have been widely used in AAD in recent years[14, 22, 23]. As a result, we also accept CNN as the backbone of the proposed DBCNet, as illustrated in Fig. 1 (b).

We employ a moving window to split EEG signals into multiple segments, which are represented by the matrix $E = [c_1, c_2, ..., c_N] \in \mathbb{R}^{T_e \times N}$, where $N = 64$ represents the total number of channels, and $c_i \in \mathbb{R}^{T_e \times 1}$ denotes the channel-wise EEG signals with the length of T_e, which varies depending on the length of the segment being analyzed. Motivated by the studies that the brain's responses to auditory stimulus could be reflected in the EEG recorded from different scalp regions [24, 25], we first assign differentiated weights to the EEG of different channels. To achieve this, we employ a learnable parameter for time domain EEG E^T and the corresponding amplitude spectrum E^F:

$$E^F = ||FFT(hann(E^T))||_2 \tag{4}$$

$$(E^T_w, E^F_w) = (E^T, E^F) \cdot w_c \tag{5}$$

where $w_c \in \mathbb{R}^N$ are the learnable channel weights and $hann$ refers to the hanning window. Here E^T and E^F shares the same shape of $1 \times T_E \times N$. Since EEG are measured in a non-invasive way, the signals suffer significant degradation duaring propagation. To enhance the EEG in each channel, we first employ $Conv$ blocks with small kernels to model it and the EEG of adjacent channels together:

$$E^{T/F}_{c1} = SELU(BN(Conv(E^{T/F}_w))) \tag{6}$$

where $E^{T/F}_{c1} \in \mathbb{R}^{4 \times T_e \times N}$ is the enhanced EEG signals and each $Conv$ block consists of a 2-D convolutional layer, a BN (batch normalization [26]) layer and a $SELU$ [27] activation function.

To exchange information between the two domains, we then employ a time-frequency interaction (TF-IA) module:

$$E^{TF} = E^T_{c1} ⓒ E^F_{c1} \tag{7}$$

$$E^{T/F}_{ia} = E^{T/F}_{c1} ⓒ SELU(BN(Conv(E^{TF}))) \tag{8}$$

The TF-IA module consists of two $Conv$ blocks and three concatenations which are denoted as ⓒ in Fig. 1(b). The two outputs of the TF-IA module have the shape of $\mathbb{R}^{C_1 \times T_e \times (N+1)}$. After the TF-IA module, a couple of $Conv$ blocks are applied to further extract speaker features. The output is a set of speaker features represented as $E^T_{c2} \in \mathbb{R}^{C_2 \times T'_e \times 1}$ and $E^F_{c2} \in \mathbb{R}^{C_2 \times T'_e \times 1}$, where C_1, C_2, T'_e equal 4, 256 and $T_e - 31$, respectively. To obtain the most representative feature, a max pooling layer is applied along the time dimension. Finally, as a tradeoff between model complexity and generalizability, FC blocks, each of which consists of a fully connection layer and a $Tanh$ activation function each, are employed to achieve a non-linear mapping as follows.

$$\hat{\mathbf{d}} = FC(FC(E^T_M) ⓒ FC(E^F_M)) \tag{9}$$

$\hat{\mathbf{d}}$ is the estimated d-vector which shares the same shape with the target d-vector $\mathbf{d}$.

3.2. Loss Function and Decision Module

The output of the DBCNet is the estimated d-vector $\hat{\mathbf{d}}$. During training, to minimize the Euclidean distance between $\hat{\mathbf{d}}$ and the attended speaker's d-vector, $\mathbf{d}_{s_a}$, while maximizing the distance between $\hat{\mathbf{d}}$ and the unattended speaker's d-vector, $\mathbf{d}_{s_u}$, the loss function is given by:

$$L = ||\hat{\mathbf{d}} - \mathbf{d}_{s_a}||_2^2 - ||\hat{\mathbf{d}} - \mathbf{d}_{s_u}||_2^2 \tag{10}$$

During the testing phase, we make decisions based on the Euclidean distance between the estimated d-vector, $\hat{\mathbf{d}}$, and the two actual d-vectors. The following principle is used to decide whether a decision y is correct:

$$y = \begin{cases} 1, & ||\hat{\mathbf{d}} - \mathbf{d}_{s_a}||_2 < ||\hat{\mathbf{d}} - \mathbf{d}_{s_u}||_2 \\ 0, & ||\hat{\mathbf{d}} - \mathbf{d}_{s_a}||_2 \geq ||\hat{\mathbf{d}} - \mathbf{d}_{s_u}||_2 \end{cases} \tag{11}$$

where 1 and 0 refer to right and wrong, respectively.

4. Experimental Setup

4.1. Data Specifications

Two publicly accessible datasets—an AAD dataset [28] and a speaker identification dataset [29], are used in our study. The AAD dataset consists of EEG recordings from 22 subjects with normal hearing who participated in 32 trials lasting 50 seconds each while listening to auditory scenes with a male and female speaker speaking simultaneously (we discard the data of one subject due to the break in the experimental session). The subjects were asked to focus on one speaker and ignore the other throughout each trial and the EEG signals were captured during the whole experiment. The details of this dataset are available in [28]. The speaker identification dataset used for d-vector extraction is detailed in [29].

For EEG, we use the same data preprocessing method as in [28], the only difference is that we employ a sampling rate of 128 Hz and a cutoff frequency of 64 Hz. The speech signals in both datasets are resampled to 16 kHz for d-vector extraction.

4.2. Training Setup

We use 6-fold cross-validation (CV) over all the trials to evaluate the proposed and the compared methods. Each subject's trials are split into six groups as evenly as feasible, one of which is utilized for testing while the others are used for training. Afterwards, EEG and speech signals are time aligned and segmented by the moving windows with 50% overlapping. Only segments where both speech streams are active are kept. Such process is repeated for six times so that segments from every trial are tested once. To remove the bias from the random grouping, the same procedure is further done five times, and the final result is the average of the five. When comparing the outcomes of different experimental configurations, we use the $Wilcoxon$-test [30] to compare the results.

For training, we use the Adam optimizer. The mini-batch size is set to 32, and the learning rate is set to 0.001. We run 80 training epochs on each model. During testing, for each EEG segment, we make decision as described in Section 3.2. The decoding accuracy is defined as the proportion of EEG segments in which auditory attention is correctly decoded. As neural network training can result in random variations from epoch to epoch, we calculate test accuracy as the median accuracy of the last five epochs as in [31].

Since there are two types of d-vectors: a short-term d-vector that corresponds to a specific speech segment and a long-term d-vector that is the centroid of all short-term d-vectors in the training dataset, we consider three training and testing strategies that use different d-vectors, as shown in Table 1. The strategy with the highest decoding accuracy is regarded as the default one.

Table 1: *D-vectors used in the three strategies and the corresponding decoding accuracy*

Strategy	Training	Testing	Accuracy (%)
I	long-term	long-term	74.2
II	long-term	short-term	76.0
III	short-term	short-term	81.3

5. Results and Analysis

5.1. Effect of Different Strategies

We first test the proposed method using the three training and testing strategies mentioned in Section 4.1. Since we focus on decoding with short windows, we conduct the experiments using a 1.0-second window, which satisfies the human attention switch criteria [23]. Across the three strategies, the proposed method achieves an average decoding accuracy of 74.2%, 76.0%, and 81.3%, respectively. The results show that the best decoding accuracy is obtained when short-term d-vectors are used in both the training and testing stages. This may be attributed to the correspondence between EEG signals and d-vectors, as variations in intonation can be reflected in both the speaker features and the listener's EEG. As a result, in the subsequent experiments, we adopt Strategy III as the default strategy.

5.2. Ablation Analysis

To demonstrate the superiority of employing speaker features in AAD and the effectiveness of the proposed DBCNet, we evaluate the decoding accuracy of three different models. The first model, called CNN-D-vector, shares the same structure as the time branch of the DBCNet. The second model, denoted as CNN-Envelope, is similar to the CNN-D-vector but replaces the max pooling layer with a LSTM layer for envelope reconstruction. The third model is the proposed DBCNet-D-vector. To control for the parameter size, we increase the channel numbers of the two CNN models such that the three models share the similar parameter size. The experiments are conducted with four different windows: 0.3-s, 0.5-s, 1.0-s, and 3.0-s, and the results are shown in Figure 2.

The CNN-Envelope achieves average decoding accuracies of 51.3% , 53.3%, 57.5%, and 68.9% for the four different windows. In contrast, the CNN-D-vector achieves significantly higher decoding ($p < 0.001$) accuracies of 73.1%, 76.2%, 77.5%, and 77.3% for the same windows, respectively. These results demonstrate the advantage of incorporating speaker features into auditory attention decoding. Specifically, the CNN-D-vector outperforms the CNN-Envelope, especially for shorter decoding windows, and the difference becomes less pronounced for longer windows. This observation supports our hypothesis that speaker features are more reliably distinguishable across different durations, while longer windows are required

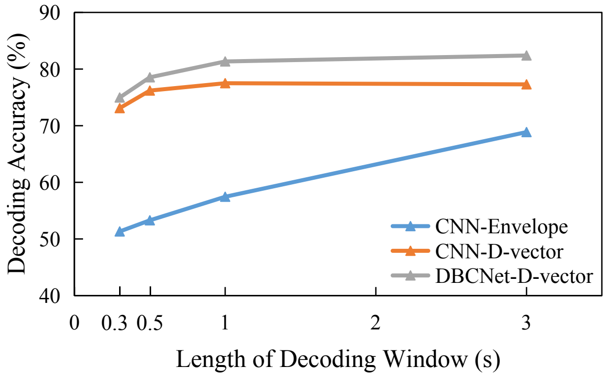

Figure 2: *Average decoding accuracy of CNN-D-vector, CNN-Envelope and DBCNet-D-vector for different window sizes over all subjects.*

Model	CCA	CRNN	DBCNet
Accuracy (%)	52.6	65.3	81.3

Table 2: *Average decoding accuracy (%) of proposed DBCNet and three other models for 1.0-s window.*

for envelope-based methods to achieve similar levels of distinguishability.

With the addition of the frequency branch to the CNN-D-vector, the DBCNet-D-vector exhibits superior performance for all windows, achieving an average accuracy of 74.9% for 0.3-s decoding window, 78.5% for 0.5-s decoding window, 81.3% for 1.0-s decoding window, and 82.4% for 3.0-s decoding window. Notably, the improvement in decoding accuracy is statistically significant ($p < 0.005$) for our focus duration 1.0-s. We attribute the improvement to the incorporation of EEG spectra, which allows for the identification and exclusion of components unrelated to auditory attention in the frequency domain. As different brain activities may occur simultaneously, the use of EEG spectra can aid in disentangling these activities and improving the accuracy of auditory attention decoding.

5.3. DBCNet vs Other Models

We further evaluate the effectiveness of the proposed method by comparing it with other decoding models. Specifically, we considere a state-of-the-art (SOTA) linear model CCA [32], a CRNN model [14] which combines the advantages of CNN and RNN. The CRNN we use has a similar parameter size with the proposed DBCNet by increasing the number and size of the convolutional layers. The window length is fixed at 1.0-s, and the results are summarized in Table 2.

The decoding accuracies of the compared models are 52.6% and 65.3%, respectively. As predicted, the proposed method achieves noticeably higher decoding accuracy ($p < 0.001$) than other models that rely on speech envelopes. These results further validate the effectiveness of the proposed method in addressing the limitations of existing approaches.

5.4. Effect of Frequency Bands

The EEG signals collected from the human brain encompass a variety of frequency bands, each of which is associated with distinct physiological activities [33]. To determine which of these frequency bands is most indicative of individual differences in speech, we investigate the decoding accuracy of the proposed method using EEG data of different frequency bands, including delta (0.1-3 Hz), theta (3-8 Hz), alpha (8-13 Hz), beta (13-30 Hz) and gamma (>30 Hz). As before, we use a 1.0-s window and the results are presented in Fig. 3.

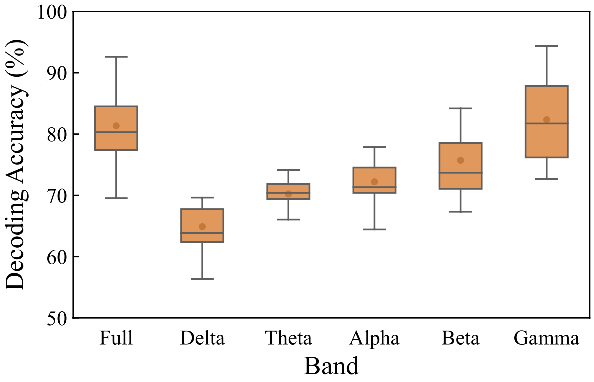

Figure 3: *Box-plot that shows the average decoding accuracy using EEG of different bands for 1.0-s window.*

Acorss the 21 subjects, the proposed model achieves average decoding accuracies of 64.9%, 70.2%, 72.2%, 75.7% and 82.3% for the the five bands. The results indicate that high decoding accuracy is achieved with high frequency bands, particularly the gamma band, with an accuracy of 82.3%. This fingding aligns with previous research [34] which established a correlation between gamma oscillations and the binding of acoustic features, such as pitch, timbre, and harmony, in speech comprehension. Interestingly, we also observe relatively low decoding accuracies with the three low frequency bands. This contrasts with the stimulus reconstruction methods, which primarily employ low frequency EEG siganls. The difference illustrates how our method differs from previous methods and the embodies the indenpendence of speech analysis and identity analysis from the perspective of decoding.

6. Conclusions

In this paper, we propose a novel AAD architecture, which creatively decodes target speaker using EEG and speaker features. To extract speaker features from EEG, we design an inventional dual branch convolutional network which takes the advantages of EEG in both time and frequency domains. To vertify the effectiveness of our architecture, we conduct several experiments on a public dataset. Results shows that our system could achieve an average decoding accuracy of $\sim$82% for 1.0-s decision window, which outperforms the methods that require speech envelope. The high performance of our architecture in time-constrained situations highlights its potential for practical applications, such as brain-informed hearing aids.

7. Acknowledgements

This research is partially supported by STI 2030—Major Projects (No. 2021ZD0201500), the National Natural Science Foundation of China (No. 12104483).

8. 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, no. 5, pp. 975–979, 1953.

[2] E. C. Bluvas and T. Q. Gentner, "Attention to natural auditory signals," *Hearing research*, vol. 305, pp. 10–18, 2013.

[3] J. R. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional gain control of ongoing cortical presentations in a "cocktail party"," *Journal of Neuroscience*, vol. 30, no. 2, pp. 620–628, 2010.

[4] S. Akram, J. Z. Simon, S. A. Shamma, and B. Babadi, "A state-space model for decoding auditory attentional modulation from meg in a competing-speaker environment," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[5] 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.

[6] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, "Real-time tracking of selective auditory attention from m/eeg: A bayesian filtering approach," *Frontiers in neuroscience*, vol. 12, p. 262, 2018.

[7] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, 2020.

[8] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, T. Francart *et al.*, "Relating eeg to continuous speech using deep neural networks: a review," *arXiv preprint arXiv:2302.01736*, 2023.

[9] N. Das, A. Bertrand, and T. Francart, "Eeg-based auditory attention decoding: Towards neuro-steered hearing devices," 2020.

[10] Z. Xu, Y. Bai, R. Zhao, Q. Zheng, G. Ni, and D. Ming, "Auditory attention decoding from eeg-based mandarin speech envelope reconstruction," *Hearing Research*, vol. 422, p. 108552, 2022.

[11] Z. Xu, Y. Bai, R. Zhao, H. Hu, G. Ni, and D. Ming, "Decoding selective auditory attention with eeg using a transformer model," *Methods*, vol. 204, pp. 410–417, 2022.

[12] 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.

[13] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of two-talker attention decoding from eeg with nonlinear neural networks and linear methods," *Scientific reports*, vol. 9, no. 1, p. 11538, 2019.

[14] Z. Fu, B. Wang, X. Wu, and J. Chen, "Auditory attention decoding from eeg using convolutional recurrent neural network," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 970–974.

[15] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[16] P. Belin, S. Fecteau, and C. Bedard, "Thinking the voice: neural correlates of voice perception," *Trends in cognitive sciences*, vol. 8, no. 3, pp. 129–135, 2004.

[17] A. W. Young, S. Frühholz, and S. R. Schweinberger, "Face and voice perception: Understanding commonalities and differences," *Trends in Cognitive Sciences*, vol. 24, no. 5, pp. 398–410, 2020.

[18] A. D. Friederici, "Towards a neural basis of auditory sentence processing," *Trends in cognitive sciences*, vol. 6, no. 2, pp. 78–84, 2002.

[19] S. Frühholz and S. R. Schweinberger, "Nonverbal auditory communication–evidence for integrated neural systems for voice signal production and perception," *Progress in Neurobiology*, vol. 199, p. 101948, 2021.

[20] S. R. Schweinberger, C. Walther, R. Zäske, and G. Kovács, "Neural correlates of adaptation to voice identity," *British Journal of psychology*, vol. 102, no. 4, pp. 748–764, 2011.

[21] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[22] M. J. Monesi, B. Accou, J. Montoya-Martinez, T. Francart, and H. Van Hamme, "An lstm based architecture to relate speech stimulus to eeg," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 941–945.

[23] 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, no. 7, pp. 2233–2242, 2022.

[24] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing the channel selection and classification accuracy in eeg-based bci," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 6, pp. 1865–1873, 2011.

[25] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel eeg: implications for online, daily-life applications," *Journal of neural engineering*, vol. 12, no. 4, p. 046007, 2015.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[27] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[28] S. A. Fuglsang, J. Märcher-Rørsted, T. Dau, and J. Hjortkjær, "Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention," *Journal of Neuroscience*, vol. 40, no. 12, pp. 2562–2572, 2020.

[29] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[30] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*. Springer, 1992, pp. 196–202.

[31] I. Kuruvila, J. Muncke, E. Fischer, and U. Hoppe, "Extracting the auditory attention in a dual-speaker scenario from eeg using a joint cnn-lstm model," *Frontiers in Physiology*, vol. 12, p. 700655, 2021.

[32] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjaer, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[33] H. Huang, J. Zhang, L. Zhu, J. Tang, G. Lin, W. Kong, X. Lei, and L. Zhu, "Eeg-based sleep staging analysis with functional connectivity," *Sensors*, vol. 21, no. 6, p. 1988, 2021.

[34] S. Palva, J. M. Palva, Y. Shtyrov, T. Kujala, R. J. Ilmoniemi, K. Kaila, and R. Näätänen, "Distinct gamma-band evoked responses to speech and non-speech sounds in humans," *Journal of neuroscience*, vol. 22, no. 4, pp. RC211–RC211, 2002.