frontiers | Frontiers in Signal Processing

TYPE Original Research
PUBLISHED 29 January 2025
DOI 10.3389/frsip.2024.1432298

OPEN ACCESS

EDITED BY
Zbyněk Koldovský,
Technical University of Liberec, Czechia

REVIEWED BY
Alberto Bernardini,
Polytechnic University of Milan, Italy
Andreas Brendel,
Fraunhofer Insititute for Integrated Circuits (IIS), Germany

*CORRESPONDENCE
Sharon Gannot,
sharon.gannot@biu.ac.il

RECEIVED 13 May 2024
ACCEPTED 09 December 2024
PUBLISHED 29 January 2025

CITATION
Gueta R, Zion-Golumbic E, Goldberger J and Gannot S (2025) Auditory attention decoding based on neural-network for binaural beamforming applications.
Front. Signal Process. 4:1432298.
doi: 10.3389/frsip.2024.1432298

Auditory attention decoding based on neural-network for binaural beamforming applications

Roy Gueta[1], Elana Zion-Golumbic[2], Jacob Goldberger[1] and Sharon Gannot[1]*

[1]Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel, [2]Multidisciplinary Brain Research Center, Bar-Ilan University, Ramat-Gan, Israel

Individuals have the remarkable ability to differentiate between speakers and focus on a particular speaker, even amidst complex acoustic environments with multiple speakers, background noise and reverberations. This selective auditory attention, often illustrated by the cocktail party problem, has been extensively researched. With a considerable portion of the population experiencing hearing impairment and requiring hearing aids, there arises a necessity to separate and decode auditory signals artificially. The linearly constrained minimum variance (LCMV) beamforming design criterion has proven effective in isolating the desired source by steering a beam toward the target speaker while creating a null toward the interfering source. Preserving the binaural cues, e.g., interaural time difference (ITFD) and interaural level difference (ILD), is a prerequisite for producing a beamformer output suitable for hearing aid applications. For that, the binaural linearly constrained minimum variance (BLCMV) beamformer generates two outputs that satisfy the standard LCMV criterion while preserving the binaural cues between the left-ear and right-ear outputs. Identifying the attended speaker from the separated speakers and distinguishing it from the unattended speaker poses a fundamental challenge in the beamformer design. Several studies showed the ability to encode essential features of the attended speech from the cortex neural response, as recorded by the electroencephalography (EEG) signals. This led to the development of several algorithms addressing the auditory attention decoder (AAD) task. This paper investigates two neural network architectures for the AAD task. The first architecture leverages transfer learning. It is evaluated using both same-trial and cross-trial experiments. The second architecture employs an attention mechanism between the speech signal represented in the short time Fourier transform (STFT) domain and a multi-band filtered EEG signal. With the goal of alleviating the problem of same-trial overfitting, this architecture employs a new data organization structure that presents the neural network (NN) with a single speaker's speech and the corresponding EEG signal as inputs. Finally, posterior probability post-processing is applied to the outputs of the NN to improve detection accuracy. The experimental study validates the applicability of the proposed scheme as an AAD method. Strategies for incorporating the AAD into BLCMV beamformer are discussed.

KEYWORDS

audio attention decoding, EEG signals, multi-microphone processing, binaural LCMV beamformer, neural network based AAD

1 Introduction

Humans possess the capability to distinguish between speakers and concentrate on a particular speaker, even in environments with multiple speakers, background noise and reverberation. The skill of selectively focusing on specific auditory cues in intricate acoustic settings, as exemplified by the cocktail party problem (Haykin and Chen, 2005; Cherry, 1953) has been extensively studied. Due to the prevalence of hearing impairment affecting a considerable portion of the population, which often requires the use of hearing aids, there arises a necessity to separate and decode auditory signals artificially.

The blind source separation (BSS) task is one of the fundamental tasks in speech processing, with numerous algorithms proposed to address this challenge (Vincent et al., 2018.; Makino. 2018) These algorithms find applications in various devices, including hearing aids, smartphones and virtual assistants. Array processing, specifically beamforming, is one of the main tools to process audio signals, specifically addressing the BSS task. It can handle audio signals contaminated by noise, reverberation, and interfering speakers. The beamformer's weights are set to satisfy specific criteria, for instance, a distortionless response to the desired speaker or a null towards the interfering source, while minimizing the output noise power. Additionally, beamformer design may require minimizing the white noise gain (WNG) to enhance resilience against system uncertainties, as discussed in Cox et al. (1987), Van Trees (2002), Pillai (2012).

One of the main design criteria for beamformers is the minimum variance distortionless response (MVDR) (Capon, 1969), which aims to minimize the noise variance while imposing a distortionless response towards the desired source. The LCMV criterion extends the MVDR criterion by applying a set of linear constraints, e.g., on the beampattern derivative in the desired direction (Er and Cantoni, 1983). In Markovich et al. (2009), the LCMV was adapted to the multi-speaker problem by directing a beam towards the desired source(s) and a null towards all interference sources while minimizing the noise level at the output. Preserving the binaural cues, e.g., ITFD and ILD, is a prerequisite for producing a beamformer output suitable for hearing aid applications. In Hadad et al. (2016), the BLCMV beamformer is introduced, and its performance is analyzed. The BLCMV beamformer generates two outputs that satisfy the standard LCMV criterion and, in addition, preserves the binaural cues. In our context, this entails steering a beam toward the desired source and nulls toward the interfering sources while preserving the binaural cues between the left-ear and right-ear outputs.

A primary challenge in designing beamformers is identifying the desired speaker, essentially informing the beamformer about which speaker is the target of interest. Several studies showed the ability to encode essential features of the attended speech from the cortex neural response, as recorded by the EEG signals. In Mesgarani and Chang (2012), the method of stimulus reconstruction was used to estimate the speech spectrogram of the attended speaker from a spectrogram of a mixture of speakers. EEG signals can be captured and utilized as input for an AAD to help determine the desired speaker and the interferer (O'Sullivan et al., 2015; Biesmans et al., 2017; Thwaites et al., 2016). Various attention decoding methods have been proposed and studied, including models based on linear temporal relative function (TRF). In Kuruvila et al. (2020), both least squares (LS) and linear minimum mean squared error (LMMSE) methods are applied to infer the relation between the EEG signal and the envelopes of the desired and interfering sources. Another approach Wong et al. (2018) is also using the cortical signal to reconstruct each of the two audio speakers in the scene. Then, a support vector machine (SVM) classifier is used to identify the desired and interfering signals.

Recent studies have harnessed the power of deep neural networks (DNNs) for auditory attention decoding. In Reddy Katthi and Ganapathy (2021), a deep multiway canonical correlation analysis (CCA) is proposed to remove artifacts from EEG recordings. This may improve the correlation between the EEG and the attended speaker signal, consequently resulting in enhanced decoding. In Fu et al. (2021), a convolutional reconstruction neural network is proposed. In Ciccarelli et al. (2019) several methods are proposed, the first is established by regular TRF analysis, the second estimates the TRF using neural network and the third performs a direct classification using DNN. In Monesi et al. (2021), a long short-term memory (LSTM) based model is proposed, and several input features, namely, mel-spectrogram, envelope, word embedding, and a combination thereof, are evaluated. Both linear, correlation-based, and convolutional neural network (CNN) methods are used in Accou et al. (2021), Vandecappelle et al. (2021) for a broad range of time frame lengths, varying from long to very short time frames. In Cai et al. (2020), it is proposed to apply the common spatial pattern (CSP) algorithm for enhancing the EEG signal. The enhanced EEG signal and audio speaker signals were then used as input to a CNN architecture.

Overfitting is a well-known and challenging phenomenon in deep learning. It occurs when the algorithm becomes overly tailored to the specific details of the training data, making it difficult to generalize to new, unseen data. Overfitting to particular trials, subjects, or datasets can lead to impressive but ultimately misleading results for AAD algorithms. In Cai et al. (2024), Rotaru et al. (2024), Puffay et al. (2023), these overfitting issues are discussed for EEG-based AAD or spatial encoding tasks.

Our study first delves into ways of using the outcomes of AAD to provide guidance to BLCMV beamformer regarding the attended speaker, thus informing it which speaker to extract.

We then explore two NN-based architectures for addressing the AAD task. The first architecture leverages transfer learning, adapting an image processing model to process both the temporal envelope of the speech signal and the raw EEG signal. The second, based on Li et al. (2021), employs an attention mechanism to analyze the speech signal represented in the STFT domain, along with multi-band filtered EEG signals.

Using two common databases Das et al. (2020) and Kuruvila et al. (2021), we present an assessment of the ability of these schemes to generalize to unseen data. Specifically, we will show that the transfer learning approach performs very well if the same trials are present during training and test stages but fails to generalize to cross-trial experiments. Moreover, we present the attention-based approach, which employs a new data organization scheme to alleviate same-trial overfitting. Nevertheless, this approach is still shown to perform very well in the same-trial case but not in the cross-trial case.

Our main contributions are: 1) applying domain adaptation for repurposing ResNet, a widely recognized DNN architecture, to the AAD task; 2) addressing the cross-trial scenario and testing the

performance of the proposed algorithms on relevant scenarios; 3) the development of a new DNN architecture for the attention decoding task; 4) presenting a new data organization paradigm for DNN training, as well as a new postprocessing method, suitable to the new organization; 5) presenting a conceptual system, integrating binaural beamformer for source separation and AAD, that may apply to speaker mixtures.

2 Problem formulation

We consider a scenario involving a human subject exposed to two separate sources, namely, the first source and the second source $\mathbf{s}_1(t,k)$ and $\mathbf{s}_2(t,k)$. Throughout the paper, we assume that the original signals are available to the AAD procedure. We only discuss potential separation schemes in Section 6 In parallel, the subject's EEG multichannel signal, $\mathbf{e}(\mathbf{t}) = e_1(t), e_2(t), \ldots, e_{C_E}(t)$, with C_E the number of EEG channels is recorded. We also define the matrix comprising all EEG signals, $\mathbf{E} = \{\mathbf{e}(\mathbf{t})\}_{t=0}^{T-1}$, where T refers to the length of the signal. We are now addressing a classification problem where, given the EEG signal and the speech signal(s), our goal is to identify the attended speaker and distinguish it from the unattended speaker. We will propose in this paper two AAD approaches and analyze their performance, advantages, and drawbacks.

3 Transfer learning approach

3.1 ResNet

We first explore a classification approach which adopts the transfer learning paradigm. In this method, we repurpose a DNN architecture originally developed for a different task and apply it to our AAD task.

ResNet, a widely recognized and extensively used DNN architecture based on residual networks (Targ et al., 2016), serves as the backbone of this method due to its proven success in image classification and various other image processing tasks. Using the ResNet architecture necessitates an additional fully connected layer to adjust the output shape to the required format:

$$\mathbf{p}_{\text{resnet}} \in \mathbb{R}^{1\times 2}. \tag{1}$$

In Equation 1, the first entry indicates the probability that the first speaker is the desired speaker, while the second entry corresponds to the probability that the second speaker is the desired speaker.

3.2 Implementation

The input to the ResNet consists of both audio speaker envelopes and the cortical signal $\mathbf{M} = [\mathbf{s}_{\text{env}_2}, \mathbf{E}, \mathbf{s}_{\text{env}_1}]^\top$,

$$\mathbf{M} \in \mathbb{R}^{(C_E+2)\times T} \tag{2}$$

where $\mathbf{s}_{\text{env}_1}$ and $\mathbf{s}_{\text{env}_2}$ represent the speech stimulus envelopes of the first and second speakers, respectively. The audio envelopes are obtained by computing the absolute value of the Hilbert transform.

Since the ResNet was originally designed for images in a 3-channel format (RGB), we concatenated three identical matrices from (Equation 2) in a tensor form: $\mathcal{M} = [\mathbf{M}, \mathbf{M}, \mathbf{M}]$, where $\mathcal{M} \in \mathbb{R}^{(C_E+2)\times T\times 3}$. Time frames of 0.5,1,2,3,4.5 s are examined, and both the KUL and FAU datasets are employed for the evaluation.

3.3 Cross-trial training and testing

To ensure an effective AAD mechanism, it is required that the employed approach will be able to generalize from the training data to real-world applications. The EEG signal inherently exhibits dynamic characteristics, showing temporal variations. Additionally, typical datasets consist of multiple trials for each subject, often with intervals separating them. This structure introduces variability between trials, suggesting distinct data patterns within each. Therefore, we propose to evaluate the generalization ability of the proposed network across trials.

We split the trials in the datasets into two phases: training/validation and testing. This partitioning enables us to evaluate our architecture's generalization ability in both the same-trial and cross-trial cases. This will allow us to detect potential overfitting tendencies by training the DNN on a subset of trials and assessing its performance on unseen trials, thus enabling us to ascertain whether the DNN is memorizing specific trial characteristics or effectively learning the underlying relationships between auditory and cortical signals.

4 Multi-band attention neural network

In this section, our attention-based neural network will be presented. The architecture, adopted from Li et al. (2021), takes speech features extracted from the STFT representation of the speech signals and the EEG multichannel signal, filtered out to several frequency bands as inputs. It then utilizes multiple fully connected (FC) layers and a cross-attention mechanism to learn the relationship between the EEG and speech inputs. This architecture will be henceforth referred to as multi-band attention neural network (MB Att. NN).

4.1 Preprocessing

The EEG data (designated as $\mathbf{E}$) was first downsampled to 128 Hz and then filtered in the range 1–50 Hz to filter out irrelevant frequency content. Five frequency bands are then considered as input features (Buzsaki and Draguhn, 2004; Viswanathan et al., 2019):

$$\text{EEG bands} = \begin{cases} \delta: & 1-3\ \text{Hz} \\ \theta: & 3-7\ \text{Hz} \\ \alpha: & 7-15\ \text{Hz} \\ \beta: & 13-30\ \text{Hz} \\ \gamma: & 30-50\ \text{Hz} \end{cases} \tag{3}$$

The EEG signal filtered into five channels will be denoted $\mathcal{E}$. To investigate the mechanisms of speech signal processing within the

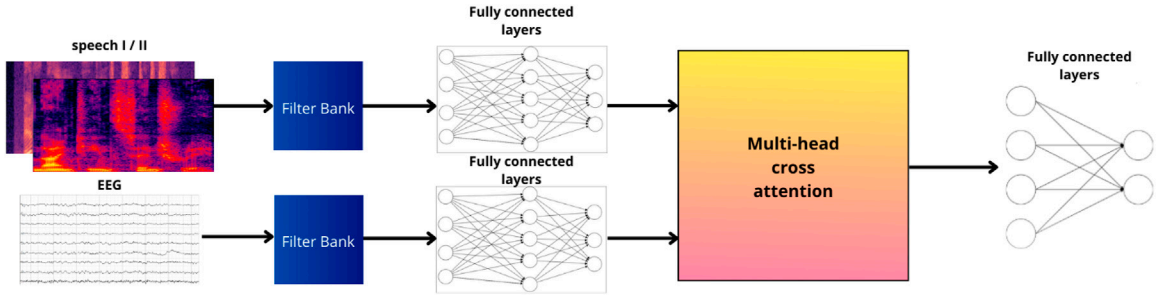

FIGURE 1
The multi-band attention neural network (MB Att. NN) architecture involves using the Gammatone spectrogram of one of the speakers as audio features and the filterbank channels defined in (Equation 3) as EEG features. These signals are embedded separately using FC layers. Subsequently, a cross-attention module is applied to both embedded features and fed into additional FC layers. The two outputs represent the probability of the processed audio signal belonging to the attended speaker.

human brain, this study employs Gammatone filtering on the audio signals, inspired by Park and Yoo (2020). Gammatone filters emulate the human auditory system's response to sound, characterized by a larger sensitivity to lower frequencies compared to higher ones. The filtering is applied in the time-frequency domain using a window size of $N_{\text{fft}} = 2048$ and an overlap of 345 samples.

4.2 Neural architecture

The top-level architecture of the NN is depicted in Figure 1. The NN has two different input features: 1) the Gammatone spectrogram auditory signal $\mathbf{S} \in \mathbb{R}^{M_S \times T}$, with T referring to the number of samples, and $M_S = 10$ to the number of Gammatone band; and 2) the cortical signal $\mathcal{E} \in \mathbb{R}^{C_E \times M_E \times T}$, where C_E refers to the number of EEG channels and $M_E = 5$ refers to the number of frequency bands.

The proposed NN architecture comprises several stages. Since the idea is to extract the relations between the audio and the EEG signals in a latent space, the network will embed both inputs separately using FC layers:

$$\begin{aligned} \mathbf{E}' &= g_E(\mathcal{E}) \in \mathbb{R}^{M_E \times C} \\ \mathbf{S}' &= g_S(\mathbf{S}) \in \mathbb{R}^{M_S \times C}, \end{aligned} \tag{4}$$

Where $g_E(\cdot)$ refers to the embedding layers for the EEG signal, comprising three FC layers with GELU and Relu activation functions, and $g_S(\cdot)$ refers to the embedding layers for the audio, comprising two FC layers with GELU and Relu activation functions, and C is the embedding size of both the EEG and audio signals. The next step is to duplicate $\mathbf{E}'$ defined in (Equation 4), namely, $\mathbf{E}' = [(\mathbf{E}')^\top, (\mathbf{E}')^\top]^\top$, in order to fit the number of frequency channels for all inputs (since $M_E = 5$ and $M_S = 10$). Then, a multi-head cross-attention is applied. Let Q be the query, K the key, and V the value. In our case, $Q = K = \mathbf{E}'$ and $V = \mathbf{S}'$. The cross-attention output represents the connection between the embedded audio and EEG signals:

$$\mathbf{A} = \text{MCA}(\mathbf{E}', \mathbf{E}', \mathbf{S}') \in \mathbb{R}^C, \tag{5}$$

where MCA refers to a multi-head cross attention layer. Finally, two additional FC layers, the second using a Sigmoid activation function, are applied to extract the probability of s_1 and s_2 to be the attended speaker:

$$\mathbf{p} = \text{FC}_2(\text{FC}_1(\mathbf{A})) \in \mathbb{R}^{1\times 2}, \tag{6}$$

where $\mathbf{p} = [p_1, p_2]$ with $p_1 = p(\#1|\mathbf{E}, s_1)$ referring to the probability that the first speaker is the attended speaker, given the first speaker signal s_1 and the EEG signal, and $p_2 = p(\#1|\mathbf{E}, s_2)$ referring to the probability that the first speaker is the attended speaker, given the second speaker signal s_2 and the EEG signal, respectively.

4.3 Unique data organization

Overfitting is a prevalent challenge in neural networks, specifically in the AAD domain (Puffay et al., 2023). The network tends to learn patterns specific to the training set and struggles to generalize to unseen data.

In many current AAD approaches, the data is structured by providing two speech signals alongside the cortical signal EEG, and with the corresponding labels (e.g., '0' for the first speaker, '1' for the second), as an input to the NN. However, a significant drawback of such a method stems from the nature of the datasets commonly utilized in this field. Each trial in these datasets is typically associated with a single desired speaker, leading to a fixed label throughout the entire trial. Since timeframes from the same trial often exist both in train and test sets, this setup allows the neural network to capitalize on trial-specific patterns, potentially learning to recognize the individual trials rather than focusing on the relationship between speech and EEG signals.

We propose a novel data organization approach to overcome the challenge of overfitting to trial-specific patterns. During training, validation, and testing, each EEG frame is presented to the neural network twice: once alongside the audio signal of the attended speaker, labeled as '1' to indicate the presence of the desired speaker, and then again with the unattended speaker's audio signal, labeled as '0'. By adopting this strategy, even when using frames from the same trial for training and testing, we alleviate the risk of the model overfitting to specific EEG trial patterns. This dissociation of the correct label from the unique characteristics of a particular trial may help enhance the model's generalization ability in the same-trial cases.

4.4 Implementation

As described, our architecture receives the preprocessed audio data represented in the time-frequency domain as input, denoted as $\mathbf{S} \in \mathbb{R}^{M_S \times T}$. Additionally, the architecture incorporates the EEG signals filtered into frequency bands, namely, $\mathcal{E} \in \mathbb{R}^{C_E \times M_E \times T}$. The effectiveness of the proposed architecture was evaluated only for window lengths of 0.5, 1, and 2 s. Window lengths of 3–5 s were not examined for this method. Two datasets, KUL and FAU (details provided subsequently), were employed for the evaluation.

4.5 Posterior probability

As shown in (Equation 6) the proposed network consists of two outputs, each representing the probability that the audio stimulus belongs to the attended speaker, as determined by the neural response encoded in the EEG signal. Let $p_1 = p(\#1|\mathbf{E}, s_1)$ denote the probability that speaker #1 is the attended speaker, and $p_2 = p(\#2|\mathbf{E}, s_2)$ denote the probability that speaker #2 is the attended speaker. We can further enhance the reliability of identifying the attended speaker by jointly utilizing the two uttered audio signals in the scene and their corresponding EEG signals. To achieve this, we calculate the posterior probability using the two audio stimuli:

$$p(\#1|\mathbf{E}, s_1, s_2) = 1 - p(\#2|\mathbf{E}, s_1, s_2) = \frac{p_1(1-p_2)}{p_1(1-p_2) + p_2(1-p_1)}. \tag{7}$$

4.6 Comparison with Li et al.

As mentioned, our architecture builds upon the work presented in Li et al. (2021). However, our approach incorporates several essential modifications compared to the original work. Firstly, in contrast to using a simple dot product for attention weights, we leverage a dedicated cross-attention layer equipped with learnable parameters, as detailed in (Equation 1). Secondly, we utilize a more suitable frontend for the EEG signal, using a 5-band filterbank. Research on EEG signals showed a relation between these specific frequency bands, especially the Alpha band, and auditory attention decoding as perceived in the human brain. Such preprocessing leverages this knowledge for better attention decoding and mitigates irrelevant information that may deteriorate performance.

Thirdly, and most importantly, we deviate from the original work by processing only a single audio input (i.e., single speaker) per iteration rather than dual audio inputs as in the original work This modification necessitates adjustments to the neural network architecture to ensure compatibility between the input shapes. Finally, due to our distinct data organization strategy, we employ posterior probability post-processing, as detailed above.

5 Experimental study

5.1 Datasets

To train and test our proposed algorithms, we used two commonly-used datasets:

5.1.1 KUL dataset

This database comprises sixteen normal hearing subjects. For each subject, a 64-channel EEG signal was acquired and sampled to the computer at 8,196 Hz. Each subject was exposed to two simultaneous speakers and instructed to focus on one while ignoring the other. The stimuli included two conditions: 1) Audio filtered with head related transfer function (HRTF) to emulate sound from 90° to the left and to the right of the head; and 2) Dichotic presentation, where two speakers were played simultaneously using earphones, one on each side. This condition does not have acoustic reflections and is hence denoted "dry." Each subject performed 20 trials. The first eight are regular trials, and the other twelve comprise partial repetitions of the former. We have, therefore, decided to use only the first eight trials in our evaluation. Further details about the dataset can be found in Das et al. (2020).

5.1.2 FAU dataset

Collected from 27 subjects, all native German speakers, the dataset comprises recordings of individuals exposed to two speech stimuli simultaneously. Participants were instructed to attend to one stimulus while ignoring the other. Each subject completed six trials, each approximately 5 min long, resulting in a total of 30 min of data per subject. The EEG device consisted of 21 channels, sampled at 2.5 kHz. Further details about this dataset can be found in Kuruvila et al. (2021).

This study will utilize the original clean audio signals for all experiments, deferring the exploration of the BLCMV beamformer outputs to future investigations.

5.2 Results

5.2.1 Competing methods

Various methods for AAD using NN based architectures have been proposed, as listed in the Introduction. To assess the performance of our proposed models, we compare them against previously published methods, as detailed in Table 1 for the KUL database.

All models considered here share a common input format: they receive two simultaneous speech stimuli and produce a single label. However, the specific data arrangement methodology employed by these works, particularly the cross-trial scenario, remains unclear and is not explicitly mentioned in the relevant papers.

5.2.2 Transfer learning approach

Here, we present the experimental results of our transfer learning-based NN architecture.

TABLE 1 Auditory attention detection accuracy (%) of three NN-based architectures for the KUL dataset.

	Time frame (s)		
	0.5	1	2
Cai et al. (2022)	-	83.6	86.9
Su et al. (2021)	84.3	86.5	88.3
Pallenberg et al. (2023)	92.8	92.8	93.0

TABLE 2 Auditory attention detection accuracy (%) using pre-trained ResNet for various time-frame sizes.

	Time frame (s)					
	0.5	1	2	3	4	5
KUL Dataset	87.1	88.2	82.4	88.7	87.9	88.3
FAU Dataset	90.5	90.9	91.9	92.6	91.2	94.5

TABLE 3 Auditory attention detection accuracy (%) using pre-trained ResNet for cross-trial approach, over different time-frame sizes.

	Time frame (s)					
	0.5	1	2	3	4	5
KUL Dataset	50.7	45.7	52.5	45.1	51.8	51.3
FAU Dataset	54.1	53.7	48.5	34.3	49.5	48.1

As shown in Table 2, the results obtained using ResNet are impressive, highlighting the efficacy of ResNet for classification tasks even with different types of data. However, despite the excellent results, verifying whether ResNet can generalize the EEG-audio relationship to unseen examples is essential. As mentioned earlier, since each subject participates in multiple trials with breaks between them, and given the rapid changes in the EEG signal, testing ResNet on examples from unseen trials is crucial to validate the reliability of the NN.

As depicted in Table 3, the results obtained in the cross-trial data arrangement show a significant degradation compared to the original data arrangement (with different segments of the same trial appearing in both training and test phases). This suggests that the success of the learning process in the previous data arrangement may have been attributed to trial overfitting rather than a genuine connection between the audio and EEG signals. Instead, it appears to detect specific patterns within each EEG trial. Furthermore, since the datasets are structured such that each trial corresponds to a specific label, the NN can classify without necessarily considering the auditory stimuli.

To further elucidate this problem, we evaluated the proposed NN performance on a unique KUL data labeling scheme. In this experiment, each trial was labeled with a distinct number. The high accuracy achieved by the scheme in this case (exceeding 90% for all trials) suggests significant overfitting. This implies that the NN 'recognizes' the specific trials based on their unique EEG patterns and corresponding labels rather than the relations between the EEG and the audio stimuli.

TABLE 4 Auditory attention detection accuracy (%) using the proposed MB Att. NN, compared for different time-frame lengths using the KUL dataset.

	Time frame (s)		
	0.5	1	2
Linear: O'Sullivan et al. (2015)	—	58.1	61.3
Linear: Geirnaert et al. (2020)	—	80.0	—
MB Att. NN	76.8	85.4	92.7

TABLE 5 Auditory attention detection accuracy (%) using the proposed multi-band attention NN, compared for different time-frame length, using FAU dataset.

	Time frame (s)		
	0.5	1	2
Linear: Kuruvila et al. (2021)	—	—	79.8
MB Att. NN	81.7	86.9	82.4

5.2.3 Multi-band attention neural network (MB Att. NN)

This subsection presents the results of the MB Att. NN. We will compare our findings to linear-based algorithms (O'Sullivan et al., 2015; Geirnaert et al., 2020; Kuruvila et al., 2021,) as these methods are not based on learning mechanisms and hence do not suffer from trial overfitting.

As depicted in Tables 4 and 5, the neural network demonstrates the ability to decode the attended speaker for both datasets accurately. As anticipated, accuracy decreases for shorter time frames, particularly evident in the KUL dataset, as the network's ability to focus on shorter patterns becomes more challenging. By analyzing the results obtained for both databases, it is evident that our proposed MB Att. NN NN model outperforms the linear methods. This is consistent with the known limitations of the latter models in capturing the complexities of short-frame data.

We stress that these results were obtained using the new and challenging data organization in which only a single speaker was available to the network during training, potentially circumventing the same-trial overfitting.

The comparison between the results obtained by baseline NN-based methods in Table 1 and the results achieved by the proposed algorithm in Table 4 reveals noteworthy findings. Despite being evaluated in a challenging experimental setup, the performance of the proposed MB Att. NN is comparable to that of (Cai et al., 2022; Su et al., 2021) and slightly inferior to that of (Pallenberg et al., 2023).

6 Informed source extraction system

AAD algorithms require two separate audio sources, as proposed in this manuscript; however, in real-life scenarios, we

typically only have access to a mixture of these sources. Therefore, it is necessary to first apply a speaker separation algorithm, preferably one that preserves spatial information, particularly binaural cues, in the context of hearing aids.

In this section, we provide a system perspective in which the AAD algorithm is incorporated into a beamformer design to determine the attended speaker, directing only this speaker to the ears of the hearing aid user. We propose utilizing the BLCMV beamformer (Hadad et al., 2012; Hadad et al., 2016) as the backbone for the separation algorithm, capable of separating the two sources while preserving their binaural cues.

6.1 Signal model

We discuss the problem of binaural hearing aids mounted on the left and right ears. Each device is equipped with a microphone array, featuring M_L microphones in the left ear and M_R microphones in the right ear. Therefore, the total number of microphones is $M = M_L + M_R$. We can formulate the scenario in the STFT domain:

$$\mathbf{z}(t,k) = \mathbf{s}_1(t,k) + \mathbf{s}_2(t,k) + \mathbf{n}(t,k), \tag{8}$$

with t the time frame index and k the frequency index. The vector of received microphone signals is denoted $\mathbf{z}(t,k)$, $\mathbf{s}_1(t,k)$ and $\mathbf{s}_2(t,k)$ are the two sources of interest in the scene as received by the microphones, and $\mathbf{n}(t,k)$ is the additive noise. Define

$$\mathbf{a}(k) = \left[a_{L,1}(k) \dots a_{L,M_L}(k), a_{R,1}(k), \dots, a_{R,M_R}(k)\right]^\top$$
$$\mathbf{b}(k) = \left[b_{L,1}(k) \dots b_{L,M_L}(k), b_{R,1}(k), \dots, b_{R,M_R}(k)\right]^\top$$

the acoustic transfer functions (ATFs) between the uttered signals $s_1(t,k)$, $s_2(t,k)$, and the microphone array, respectively. Then we can rewrite (Equation 8) as:

$$\mathbf{z}(t,k) = \mathbf{a}(k)s_1(t,k) + \mathbf{b}(k)s_2(t,k) + \mathbf{n}(t,k), \tag{9}$$

where

$$\mathbf{z}(t,k) = \left[z_{L,1}(t,k) \dots z_{L,M_L}(t,k), z_{R,1}(t,k), \dots, z_{R,M_R}(t,k)\right]^\top$$

is the measurement vector comprising all the left and right microphone signals. We assume the ATFs, $\mathbf{a}(k)$ and $\mathbf{b}(k)$, to be time-invariant, or at most slowly-time varying, and therefore omit the time index. For brevity, we will omit the time and frequency index hereinafter.

To facilitate the development of the binaural beamformer, we reformulate (Equation 9) w.r.t. to the parameters of the left and right arrays. Let M_r, M_l be the indexes of the reference microphones on the right and left arrays, respectively, and a_{M_l}, a_{M_r}, b_{M_l}, b_{M_r} the corresponding ATFs. Thus, the measurement vector in (11) can be reexpressed in terms of both the left and right reference microphones.

$$\mathbf{z} = \frac{\mathbf{a}}{a_{M_l}}\left(s_1 a_{M_l}\right) + \frac{\mathbf{b}}{b_{M_l}}\left(s_2 b_{M_l}\right) + \mathbf{n} = \frac{\mathbf{a}}{a_{M_r}}\left(s_1 a_{M_r}\right) + \frac{\mathbf{b}}{b_{M_r}}\left(s_2 b_{M_r}\right) + \mathbf{n}.$$

We are now ready to define the relative transfer functions (RTFs) as the ATF normalized by the left and right reference microphone, respectively:

$$\tilde{\mathbf{a}}_l = \frac{\mathbf{a}}{a_{M_l}}, \quad \tilde{\mathbf{a}}_r = \frac{\mathbf{a}}{a_{M_r}}, \quad \tilde{\mathbf{b}}_l = \frac{\mathbf{b}}{b_{M_l}}, \quad \tilde{\mathbf{b}}_r = \frac{\mathbf{b}}{b_{M_r}},$$

and the respective normalized signals:

$$\tilde{s}_{1,l} = s_1 a_{M_l}, \quad \tilde{s}_{1,r} = s_1 a_{M_r}, \quad \tilde{s}_{2,l} = s_2 a_{M_l}, \quad \tilde{s}_{2,r} = s_2 a_{M_r}.$$

Finally, the measured microphone signals can be explicitly rewritten in terms of the left and right RTFs:

$$\mathbf{z} = \tilde{\mathbf{a}}_l \tilde{s}_{1,l} + \tilde{\mathbf{b}}_l \tilde{s}_{2,l} + \mathbf{n} = \tilde{\mathbf{a}}_r \tilde{s}_{2,r} + \tilde{\mathbf{b}}_r \tilde{s}_{2,r} + \mathbf{n}. \tag{10}$$

We stress that in (Equation 10), the same measurement vector $\mathbf{z}(t,k)$ is written in terms of both the left and right RTFs, thus enabling the development of the BLCMV beamformer that preserves the binaural cues of the desired and interference sources (Hadad et al., 2012; Hadad et al., 2016).

6.2 The binaural linearly constrained minimum variance (BLCMV) beamformer

The BLCMV beamformer extends the regular LCMV beamformer (Markovich et al., 2009) to the binaural case. It is designed with two sets of linear constraints. The first constraint, applied to the desired source, is a distortionless response:

$$\mathbf{w}_l^{\mathrm{H}} \tilde{\mathbf{a}}_l = 1, \quad \mathbf{w}_r^{\mathrm{H}} \tilde{\mathbf{a}}_r = 1. \tag{11}$$

The second constraint set, applied to the interferer source, is responsible for its attenuation:

$$\mathbf{w}_l^{\mathrm{H}} \tilde{\mathbf{b}}_l = \eta, \mathbf{w}_r^{\mathrm{H}} \tilde{\mathbf{b}}_r = \eta \tag{12}$$

with η the *attenuation factor* of the interferer satisfying $0 < \eta \ll 1$. Applying this dual constraint set also guarantees the preservation of the binaural cues of both the desired and interference sources when introduced to both hearing aid devices. Based on the constraints on (Equations 11, 12), the left and right RTF constraint matrices can now be defined:

$$\tilde{\mathbf{C}}_l = \begin{bmatrix} \tilde{\mathbf{a}}_l & \tilde{\mathbf{b}}_l \end{bmatrix}, \quad \tilde{\mathbf{C}}_r = \begin{bmatrix} \tilde{\mathbf{a}}_r & \tilde{\mathbf{b}}_r \end{bmatrix},$$

and the corresponding response vectors for extracting source #1 are given by:

$$\mathbf{g}_l^1 = \begin{bmatrix} 1, & \eta \end{bmatrix}^\top, \quad \mathbf{g}_r^1 = \begin{bmatrix} 1, & \eta \end{bmatrix}^\top. \tag{13}$$

Alternatively, we can define the response vectors for extracting s_2 rather than s_1 as:

$$\mathbf{g}_l^2 = \begin{bmatrix} \eta, & 1 \end{bmatrix}^\top, \quad \mathbf{g}_r^2 = \begin{bmatrix} \eta, & 1 \end{bmatrix}^\top. \tag{14}$$

The left and right BLCMV weights are given by following constrained optimization using either the left or right response vector:

$$\mathbf{w}_l^{1/2} = \underset{\mathbf{w}_l}{\operatorname{argmin}} \{\mathbf{w}_l^{\mathrm{H}} \mathbf{R}_N \mathbf{w}_l\} \quad \text{s.t. } \tilde{\mathbf{C}}_l^{\mathrm{H}} \mathbf{w}_l = \mathbf{g}_l^{1/2},$$
$$\mathbf{w}_r^{1/2} = \underset{\mathbf{w}_r}{\operatorname{argmin}} \{\mathbf{w}_r^{\mathrm{H}} \mathbf{R}_N \mathbf{w}_r\} \quad \text{s.t. } \tilde{\mathbf{C}}_r^{\mathrm{H}} \mathbf{w}_r = \mathbf{g}_r^{1/2}$$

with $\mathbf{R}_N = E\{\mathbf{n}\mathbf{n}^{\mathrm{H}}\}$. The closed-form solution of the BLCMV beamformer is given by Hadad et al. (2016):

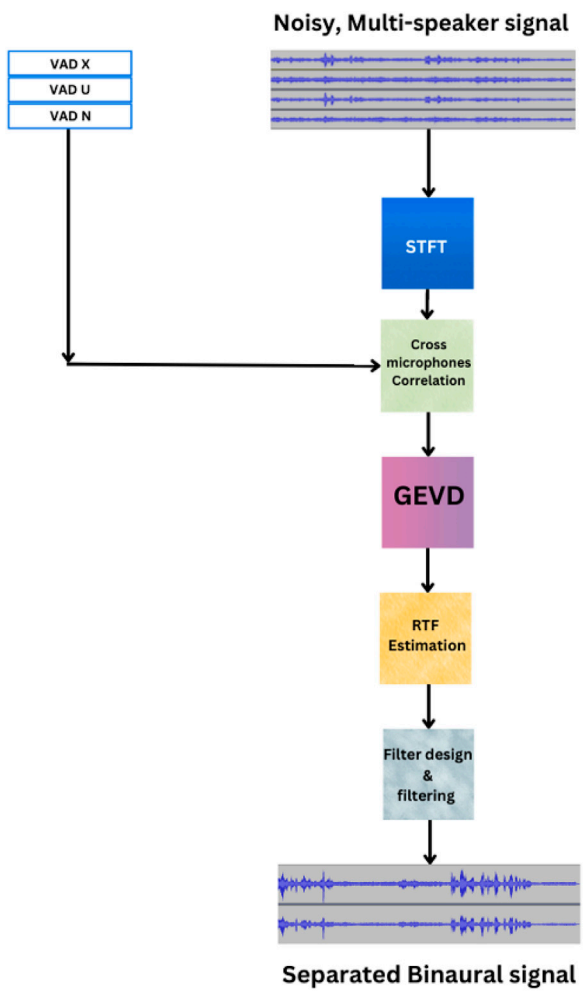

FIGURE 2
The block diagram of the BLCMV beamformer.

$$\mathbf{w}_l^{1/2} = \mathbf{R}_N^{-1}\tilde{\mathbf{C}}_l\left[\tilde{\mathbf{C}}_l^{\mathrm{H}}\mathbf{R}_N^{-1}\tilde{\mathbf{C}}_l\right]^{-1}\mathbf{g}_l^{1/2} \tag{15}$$

$$\mathbf{w}_r^{1/2} = \mathbf{R}_N^{-1}\tilde{\mathbf{C}}_r\left[\tilde{\mathbf{C}}_r^{\mathrm{H}}\mathbf{R}_N^{-1}\tilde{\mathbf{C}}_r\right]^{-1}\mathbf{g}_r^{1/2}. \tag{16}$$

The binaural outputs are then given by:

$$\hat{s}_{1,l} = \left(\mathbf{w}_l^1\right)^{\mathrm{H}}\mathbf{z}, \quad \hat{s}_{1,r} = \left(\mathbf{w}_r^1\right)^{\mathrm{H}}\mathbf{z}, \tag{17}$$

$$\hat{s}_{2,l} = \left(\mathbf{w}_l^2\right)^{\mathrm{H}}\mathbf{z}, \quad \hat{s}_{2,r} = \left(\mathbf{w}_r^2\right)^{\mathrm{H}}\mathbf{z} \tag{18}$$

Depending on the definitions of $\mathbf{g}_l$ and $\mathbf{g}_r$ (either (Equation 13) or Equation 14). The two alternative implementations extract either source s_1 or source s_2.

6.3 The system perspective

The implementation of (B)LCMV beamformer necessitates a control mechanism to determine the speakers' activity patterns. In Chazan et al. (2018), an NN-based control mechanism is proposed to determine the activities of the speakers by classifying speech segments into three classes: 1) no active speaker, 2) only one speaker is active, and 3) more than one speaker is active. During class #0, the noise correlation matrix $\mathbf{R}_N$ can be estimated; during class #1, the RTF of the active speaker is determined, and during class #2, no estimation takes place. The BLCMV beamformer is illustrated in Figure 2

While such a system is sufficient for implementing the beamformer in (Equation 15, 16), it remains to determine which of the speakers is desired and which is the interferer, i.e., to select between the two alternatives of $\mathbf{g}_l$ and $\mathbf{g}_r$, or in other words, selecting between (Equation 13) and (Equation 14). As explained below, this issue can be resolved by implementing an AAD.

In our prospective integrated BLCMV-AAD system, the AAD algorithm identifies the user's attended speaker, acting on the separated audio streams generated by the beamformer and the EEG channels.

As illustrated in Figure 3, the separated streams are fed into the AAD for identifying the attended speaker, which is then presented to the user. The input to the AAD module will be the EEG signal alongside either separated signal. The AAD will determine the signal of interest in the acoustic scene by comparing the classification results of proposed scheme (see Equation 7). The attenuation factor η can take two values. For determining the attended speaker, it is preferred to set $\eta = 0$, so that only the candidate speech signal is presented to the AAD. For rendering the acoustic scene, it is recommended to set η to a higher value, typically $\eta \approx 0.1$, to ensure that the binaural cues of the unattended speaker are preserved, although it is significantly attenuated.

In the current manuscript, we stress that only the original uttered signals are used as inputs, namely, $s_1(t), s_2(t)$. In the envisioned system, we will feed the AAD with the two alternative outputs of the BLCMV beamformer either $\hat{s}_{1,l}(t), \hat{s}_{2,l}(t)$ (or $\hat{s}_{1,r}(t), \hat{s}_{2,r}(t)$), as defined in (Equations 18, 17), and determine which of them is the attended speaker (see Section 4.5).[1]

7 Discussion and conclusion

Decoding attention in real-life scenarios remains a challenging and active research area. Although the analysis of cortical signals holds promise for addressing this challenge, practical solutions are still unavailable. Linear models have traditionally shown high decoding accuracy over long time frames. However, their performance tends to suffer in real-time scenarios, especially with short utterances, resulting in reduced reliability. In recent years, the emergence of nonlinear DNNs has opened up new avenues for tackling this issue, offering various architectures to explore.

Accurate AAD holds particular significance for beamformer guidance tasks, where the objective is to identify the user's desired speaker in multiple speaker scenarios. By decoding auditory attention from the user's EEG signal, such an algorithm can provide real-time feedback regarding the target speaker. This information can then be used to direct the beamformer (in our case, the BLCMV beamformer (Hadad et al., 2012; Hadad et al., 2016) to provide the desired audio stream to the hearing device wearer. Furthermore, such an AAD

1 Validation of such a system will require online implementation of the separation algorithm and AAD for presenting the correct speech utterance to wearer of the EEG sensors, and is left for a future implementation study.

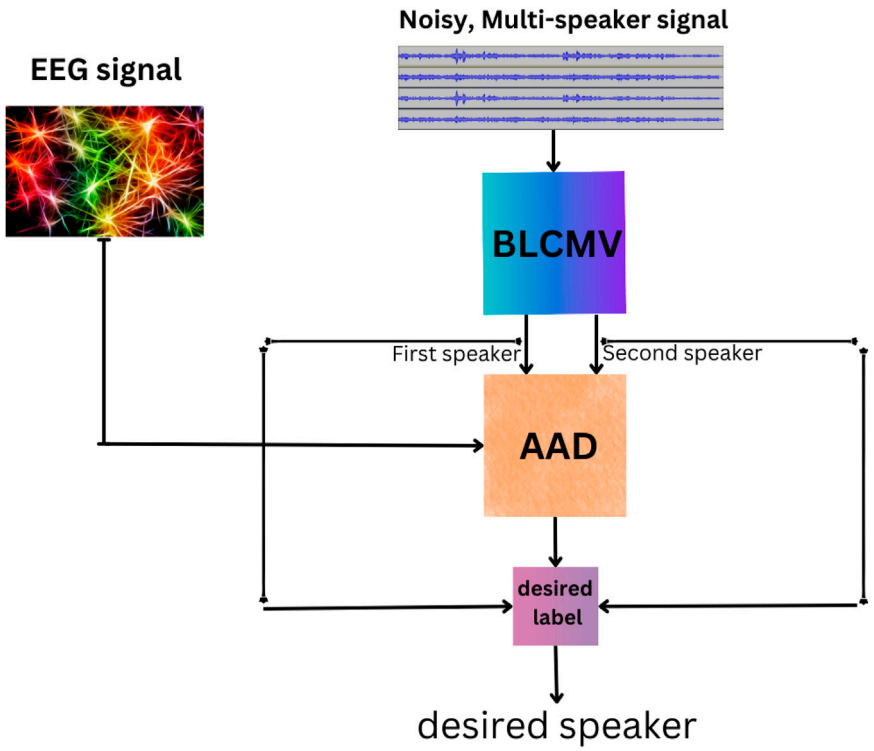

FIGURE 3
Block diagram of the proposed system with the AAD and the BLCMV beamformer.

algorithm can be integrated with the beamformer to create a closed-loop system. The AAD module could continuously identify the desired speaker, informing the beamformer to adjust the audio stream accordingly. This real-time feedback loop would facilitate the user's ability to focus on the target speaker by enhancing the desired audio stream and suppressing interfering speech. Nevertheless, in this contribution, we only used the original attended and unattended speech signal in the experimental study. Substituting these signals with the outputs of the BLCMV beamformer is left for a future study, as it requires significant implementation efforts. We still believe that the proposed system perspective elucidates the potential of AAD-controlled hearing aid algorithms.

In this paper, we have proposed two DNN-based AAD methods. The first is based on transfer learning. We propose leveraging the power of ResNet architecture, which is known for its robustness in image-processing tasks. A thorough analysis indicates that this method demonstrates favorable outcomes when segments from the same trial are available for both training and testing phases (same-trial scenario). However, performance declines significantly in the cross-trial scenario. These findings are in line with the finding in Rotaru et al. (2024), which recognizes the biases introduced by EEG-based auditory attention decoder (AAD).

We then proposed a new method, adopted from Li et al. (2021), with several critical architectural modifications, e.g., using an attention mechanism to extract the relations between the EEG and audio embeddings. Moreover, recognizing the risk of overfitting related to assigning unique labels to each EEG trial, we implemented a novel data organization strategy. Under this data structure, the desired and interfering audio signals are alternately, rather than concurrently, presented to the NN alongside the EEG signal. While this strategy effectively addresses challenges associated with the same-trial setup, it is unable to handle the cross-trial scenario.

Our AAD approaches were extensively tested using two widely-used datasets, namely, the KUL Das et al. (2020) and Kuruvila et al. (2021) datasets, and achieved advantageous results compared with baseline methods, both linear and DNN-based.

While achieving promising results, there is still room for improvement. Future work will pursue two orthogonal directions. First, we will implement and analyze the influence of substituting the original attended and unattended speakers' signals by the outputs of the BLCMV beamformer and propose methods for robustifying the overall integrated system. Next, we aim to tackle more challenging problems such as cross-trial or even cross-subject learning. This endeavor is intended to enable the seamless integration of advanced hearing aids with minimal subject-specific optimization.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

RG: Investigation, Software, Validation, Writing–original draft. EZ-G: Conceptualization, Funding acquisition, Methodology,

Resources, Supervision, Writing–review and editing. JG: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing–review and editing. SG: Conceptualization, Funding acquisition, Investigation, Methodology, Resources, Supervision, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The work was supported by the Israeli Ministry of Science and Technology. This work was supported by Deutsche Forschungsgemeinschaft (DFG) grant # 490839860 to EZ-G.

Acknowledgments

The authors express their gratitude to Mr. Pinchas Tandeitnik for his assistance in setting up the experimental setup and Mr. Eshed Rabinovitz for the fruitful discussions and his advice concerning the analysis of the results.

Conflict of interest

The authors declare that the research was conducted without any commercial or financial relationships that could potentially create a conflict of interest.

The handling editor ZK declared a past co-authorship with the author SG.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Accou, B., Monesi, M. J., Hamme, H. V., and Francart, T. (2021). Predicting speech intelligibility from EEG using a dilated convolutional network. arXiv preprint 2105.06844

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2017). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Trans. Neural Syst. Rehabilitation Eng.* 25, 402–412. doi:10.1109/tnsre.2016.2571900

Buzsaki, G., and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science* 304, 1926–1929. doi:10.1126/science.1099745

Cai, S., Su, E., Song, Y., Xie, L., and Li, H. (2020). *Low latency auditory attention detection with common spatial pattern analysis of EEG signals*. ISCA—International Speech Communication Association, 2772–2776.

Cai, S., Su, E., Xie, L., and Li, H. (2022). EEG-based auditory attention detection via frequency and channel neural attention. *IEEE Trans. Human-Machine Syst.* 52, 256–266. doi:10.1109/thms.2021.3125283

Cai, S., Zhang, R., and Li, H. (2024). "Robust decoding of the auditory attention from EEG recordings through graph convolutional networks," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2320–2324.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* 57, 1408–1418. doi:10.1109/proc.1969.7278

Chazan, S. E., Goldberger, J., and Gannot, S. (2018). "DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming," in *IEEE international conference on audio and acoustic signal processing (ICASSP)*.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25, 975–979. doi:10.1121/1.1907229

Ciccarelli, G., Nolan, M., Perricone, J., Calamia, P. T., Haro, S., O'Sullivan, J., et al. (2019). Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods. *Sci. Rep.* 9, 11538. doi:10.1038/s41598-019-47795-0

Cox, H., Zeskind, R., and Owen, M. (1987). Robust adaptive beamforming. *IEEE Trans. Acoust. Speech, Signal Process.* 35, 1365–1376. doi:10.1109/tassp.1987.1165054

Das, N., Francart, T., and Bertrand, A. (2020). *Auditory attention detection dataset KULeuven*. Zenodo.

Er, M., and Cantoni, A. (1983). Derivative constraints for broad-band element space antenna array processors. *IEEE Trans. Acoust. Speech, Signal Process.* 31, 1378–1393. doi:10.1109/tassp.1983.1164219

Fu, Z., Wang, B., Wu, X., and Chen, J. (2021). "Auditory attention decoding from EEG using convolutional recurrent neural network," in *29th European signal processing conference (EUSIPCO)*, 970–974.

Geirnaert, S., Francart, T., and Bertrand, A. (2020). Fast EEG-based decoding of the directional focus of auditory attention using common spatial patterns. *IEEE Trans. Biomed. Eng.* 68, 1557–1568. doi:10.1109/tbme.2020.3033446

Hadad, E., Doclo, S., and Gannot, S. (2016). The binaural LCMV beamformer and its performance analysis. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 24, 543–558. doi:10.1109/taslp.2016.2514496

Hadad, E., Gannot, S., and Doclo, S. (2012). "Binaural linearly constrained minimum variance beamformer for hearing aid applications," in *International workshop on acoustic signal enhancement (IWAENC)*.

Haykin, S., and Chen, Z. (2005). The cocktail party problem. *Neural Comput.* 17, 1875–1902. doi:10.1162/0899766054322964

Kuruvila, I., Demir, K. C., Fischer, E., and Hoppe, U. (2021). Inference of the selective auditory attention using sequential LMMSE estimation. *IEEE Trans. Biomed. Eng.* 68, 3501–3512. doi:10.1109/tbme.2021.3075337

Kuruvila, I., Fischer, E., and Hoppe, U. (2020). "An LMMSE-based estimation of temporal response function in auditory attention decoding," in *42nd annual international conference of the IEEE engineering in medicine biology society (EMBC)*, 2837–2840.

Li, P., Cai, S., Su, E., and Xie, L. (2021). A biologically inspired attention network for EEG-based auditory attention detection. *IEEE Signal Process. Lett.* 29, 284–288. doi:10.1109/lsp.2021.3134563

Makino, S. (2018). *Audio source separation*. Springer.

Markovich, S., Gannot, S., and Cohen, I. (2009). Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Trans. Audio, Speech, Lang. Process.* 17, 1071–1086. doi:10.1109/tasl.2009.2016395

Mesgarani, N., and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–236. doi:10.1038/nature11020

Monesi, M. J., Accou, B., Francart, T., and Van Hamme, H. (2021). Extracting different levels of speech information from EEG using an LSTM-based model. Interspeech 2021, 526–530. doi:10.21437/Interspeech.2021-336

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25, 1697–1706. doi:10.1093/cercor/bht355

Pallenberg, R., Griedelbach, A.-K., and Mertins, A. (2023). "LSTMs for EEG-based auditory attention decoding," in *European signal processing conference (EUSIPCO)*, 1055–1059.

Park, H., and Yoo, C. D. (2020). CNN-based learnable gammatone filterbank and equal-loudness normalization for environmental sound classification. *IEEE Signal Process. Lett.* 27, 411–415. doi:10.1109/lsp.2020.2975422

Pillai, S. U. (2012). *Array signal processing*. Springer Science and Business Media.

Puffay, C., Accou, B., Bollens, L., Monesi, M. J., Vanthornhout, J., Francart, T., et al. (2023). Relating EEG to continuous speech using deep neural networks: a review. *J. Neural Eng.* 20, 041003. doi:10.1088/1741-2552/ace73f

Reddy Katthi, J., and Ganapathy, S. (2021). Deep correlation analysis for audio-EEG decoding. *IEEE Trans. Neural Syst. Rehabilitation Eng.* 29, 2742–2753. doi:10.1109/tnsre.2021.3129790

Rotaru, I., Geirnaert, S., Heintz, N., Van de Ryck, I., Bertrand, A., and Francart, T. (2024). What are we really decoding? unveiling biases in EEG-based decoding of the spatial focus of auditory attention. *J. Neural Eng.* 21, 016017. doi:10.1088/1741-2552/ad2214

Su, E., Cai, S., Li, P., Xie, L., and Li, H. (2021). "Auditory attention detection with EEG channel attention," in *Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, 5804–5807.

Targ, S., Almeida, D., and Lyman, K. (2016). Resnet in resnet: generalizing residual architectures. *arXiv Prepr. arXiv:1603.08029*. doi:10.48550/arXiv.1603.08029

Thwaites, A., Glasberg, B., Nimmo-Smith, I., Marslen-Wilson, W. D., and Moore, B. C. J. (2016). Representation of instantaneous and short-term loudness in the human cortex. *Front. Neurosci.* 10, 183. doi:10.3389/fnins.2016.00183

Vandecappelle, S., Deckers, L., Das, N., Ansari, A. H., Bertrand, A., and Francart, T. (2021). EEG-based detection of the locus of auditory attention with convolutional neural networks. *eLife* 10, e56481. doi:10.7554/elife.56481

Van Trees, H. L. (2002). *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley and Sons.

Vincent, E., Virtanen, T., and Gannot, S. (2018). *Audio source separation and speech enhancement* (Wiley).

Viswanathan, V., Bharadwaj, H. M., and Shinn-Cunningham, B. G. (2019). Electroencephalographic signatures of the neural representation of speech during selective attention. *ENeuro* 6, ENEURO.0057–19.2019. doi:10.1523/eneuro.0057-19.2019

Wong, D. D. E., Fuglsang, S. A., Hjortkjær, J., Ceolini, E., Slaney, M., and de Cheveigné, A. (2018). A comparison of temporal response function estimation methods for auditory attention decoding. *Front. Neurosci.* 12, 531.