APSIPA Transactions on Signal and Information Processing, 2023, 12, e22

Overview Paper

EEG-based Auditory Attention Detection in Cocktail Party Environment

Siqi Cai[1], Hongxu Zhu[1], Tanja Schultz[2] and Haizhou Li[3,4*]

[1]*Department of Electrical and Computer Engineering, National University of Singapore, Singapore*
[2]*Cognitive Systems Lab, University of Bremen, Germany*
[3]*Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China*
[4]*Machine Listening Lab, University of Bremen, Germany*

ABSTRACT

The cocktail party effect refers to a challenging problem in speech perception where one is able to selectively attend to one sound source in a noisy and multi-talk environment. The recent studies in neuroscience and psychoacoustics shed light on how the human brain solves the cocktail party problem, that inspires many computational solutions. With the advent of novel physiological techniques and deep learning algorithms, it is now possible to effectively detect auditory attention based on brain signals. In this paper, we provide a comprehensive overview of the most recent EEG-based auditory attention detection techniques and the methods to evaluate their performance. We examine both statistical and deep learning approaches, exploring their strengths and limitations. Furthermore, we also point out the gaps between the state-of-the-art and the practical needs in real-world applications. We also offer an overview of the available resources for EEG-based auditory attention detection research.

*Corresponding author: Haizhou Li, haizhouli@cuhk.edu.cn.

Received 03 May 2023; Revised 07 July 2023
ISSN 2048-7703; DOI 10.1561/116.00000128

1 Introduction

Speech perception is a cognitive process that enables us to interpret and understand our acoustic environment. Although we often take the discrimination, identification, and interpretation of acoustic signals for granted, speech perception is a complex motor process that begins in the cochlea, travels through the auditory nerve and several auditory nuclei, and ends in the primary auditory cortex and different brain regions [91]. For individuals with normal hearing, speech perception may seem straightforward, yet the limitations of this ability are revealed in the presence of background noise, particularly among the elderly and those with hearing loss [93]. Indeed, high-intensity non-speech noise can obscure sounds and make words ambiguous or unintelligible. When the noise is composed primarily of other speakers, it may also distract the listener's attention, creating a more complicated scenario, namely the cocktail party problem [27].

The inability to follow a single speaker in a cocktail party situation is usually the first symptom of a speech perception problem for most people [90]. Such a speech perception problem is not only common in elders but also afflicts young adults with mild hearing loss and cochlear implant recipients [91]. Moreover, many people with normal hearing thresholds may experience difficulties understanding speech in noisy environments [7]. With the increasing number of people with hearing problems, it is crucial to explore the underlying mechanisms of selective listening at cocktail party scenarios for a better understanding of hearing loss, and further improve the hearing function in difficult listening conditions. Since the 1950s, the cocktail party problem has been the subject of research in a wide range of disciplines, including physiology, neurobiology, psychophysiology, cognitive psychology, biophysics, computer science, and engineering [17, 56].

In the context of a cocktail party, speech perception entails two fundamental tasks: speech separation and selective auditory attention [75]. Human ears collect a mixture of signals from all sound sources in the auditory scene. However, the listener may be interested only in one particular sound source. Hence, empowering a hearing-aid device to extract the target speech from the mixture will greatly benefit speech perception for hearing-impaired individuals. The study of computational solutions to speech separation is worthy of another full overview, which is not the focus of this article. Interested readers are referred to [112] for in-depth discussions. In this paper, we are particularly interested in the second task, that is to automatically detect the auditory attention of a listener from his/her brain signals. As the human brain is born with the auditory attention ability in the cocktail party, the findings on the brain's "magic" not only advance the understanding of related clinical studies, but also offer valuable insights into effective interventions for clinical populations who may experience challenges in speech perception.

The exploration of how the human brain solves the cocktail party problem has been a sustained effort. It is common to keep the subject in a cocktail party environment and monitor the associated neural response in his/her brain. The functional methods in these experiments mainly fall into two broad categories [98]. One is the hemodynamic measurements including functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and functional near-infrared spectroscopy (fNIRS), among others, in which we learned the neural activity in the whole brain through the changes to blood flow. For instance, Peelle and Wingfield discovered that focusing on a speech at a cocktail party environment activates more brain regions than hearing speech that is acoustically clear [91]. Another is the studies of the activity in brain neurons using various biological signals, including both invasive methods such as Electrocorticography (ECoG) and stereoelectroencephalography (sEEG), as well as non-invasive methods like electroencephalography (EEG) and magnetoencephalography (MEG). These studies revealed that biological signals respond preferentially to critical features of the attended speech (such as temporal representations [5, 39, 78, 88, 96] and spatial locations [37, 114, 116]) rather than mixture speeches.

The advancement of these neuroimaging and neurophysiological studies has greatly benefited speech perception in cocktail party scenarios as well. As the neural response of the brain is closely related to attention, it is logical to hypothesize that one can detect auditory attention from brain signals. This topic has gained increasing traction in the past decade, and is generally referred to as *auditory attention detection (AAD)*. The success of AAD opens up the possibilities of neuro-steered smart hearing devices, which detect a listener's auditory attention so as to select a sound source from a complex acoustic environment just like what humans do. A number of biological signals may carry such auditory attention trace. Not all of them are appropriate for neuro-steered hearing devices. For instance, hemodynamic measurements have a long data collection latency, the invasive EEG and ECoG may be harmful to the population outside of the clinical treatment, and MEG is not wearable. In contrast, EEG enjoys the superiority of being less expensive, more widely available, and easier to use, making it a viable option for integration into everyday devices and future brain-computer interface (BCI) applications. Therefore, we mainly limited the scope of this article to EEG-based AAD methods.

The training and run-time inference of a typical EEG-based speech perception system is illustrated in Figure 1. In the offline training phase, the model learns to associate the EEG signals and their speech stimuli. During the run-time inference, the attended speech will be determined and enhanced to improve speech perception. These techniques can be categorized in different ways, for example, according to the type of speech stimulus - clean speech vs speech mixtures; the type of attention focus - speaker vs locus, the type of EEG data - full scalp EEG vs ear-EEG; the workflow of the AAD model -

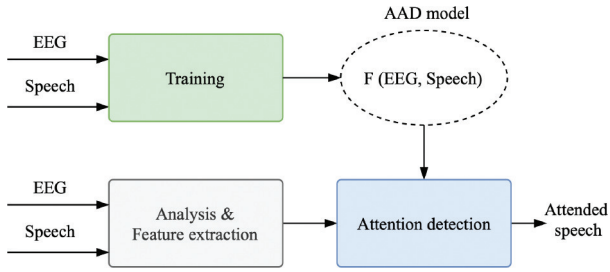

Figure 1: Typical flow of EEG-based speech perception in a cocktail party environment involves training the auditory attention detection (AAD) model, represented by the green box, and applying the detection in real-time, represented by the blue box.

stimulus-reconstruction vs direct classification; the generalization ability - subject-dependent vs subject-independent. From the viewpoint of the speech stimulus, early EEG-based speech perception research was premised on the assumption of ideal speech separation performance. These studies utilized the ground truth of speech stimulus directly as a reference. With clean speech, one can study a linear or non-linear reconstruction function to estimate the acoustic features of the attended speech, e.g. envelope, spectrogram, Mel-spectrogram, linguistic speech representations, from the full scalp EEG recordings, or vice versa. Although such reconstruction functions cannot perfectly reconstruct the stimuli, auditory attention can be determined by the correlation between the output with the ground-truth feature. This pipeline has been widely studied with different approaches, just name a few, linear regression (LS) [88], canonical correlation analysis (CCA) [28], averaging decoders [88], averaging auto-correlation matrices [12]. Later, with the advancements in deep learning, some studies aimed to reconstruct the envelope of the target speaker's signal from EEG signals using non-linear neural networks (NNs) [105] and long short-term memory (LSTM) model [82].

We posit that the stimulus-reconstruction approach exhibits two limitations. Firstly, the process of stimulus reconstruction and correlation evaluation is not optimized to effectively detect attention. Secondly, the compression of multi-channel EEG signals into a single waveform through stimulus reconstruction reduces the available information for analysis. While such transformation interprets well how brain signals correlate with speech stimulus, it doesn't necessarily represent the best way for auditory attention detection. To avoid any information loss in data compression, some recent works intended to classify the attended speaker [29] or locus [110] directly, and have achieved great success.

Since clean speech is not always available, especially in real-life scenarios, studies on EEG-based speech perception are conducted to estimate the selective attention from the mixture to improve the feasibility of the neuro-steered hearing devices. Coordination with speech separation [11, 33, 109] is one of the

research directions to overcome this challenge. Besides, in [25, 63], the attended speech is directly extracted from the mixture with the additional feature estimated from EEG data. Another direction to increase the practicability is making the system more portable, which is achieved by reducing the number of the required EEG electrodes [22, 81, 83–85, 101], using different types of EEG recording equipment, such as ear-EEG [14, 15, 36, 38, 42, 43, 61, 64–66, 73, 76, 77, 80].

Geirnaert *et al.* [49] presented an overview of EEG-based AAD, which summarizes the traditional modeling approach. With the advent of deep learning, EEG-based speech perception techniques have seen a significant advancement. The neural solution has not only enhanced the existing state-of-the-art methods but also bridged the gap between the ideal model and the practical implementation of neuro-steered hearing devices in noisy environments. Additionally, it opens up a new avenue of research beyond the existing stimulus-reconstruction and direct classification AAD methodologies. Nonetheless, these classical AAD approaches have played a critical role in advancing the understanding of speech perception in cocktail party environments and provided valuable insights into various aspects of the research challenge. With this paper, we aim to offer a comprehensive overview of EEG-based AAD research for speech perception in cocktail party scenarios by presenting a perspective that highlights the core design principles, ranging from the ideal AAD model to practical implementation, along with the challenges encountered and the future directions of the field.

This paper is organized as follows: Section 2 introduces the fundamentals of how the brain's auditory system perceives speech. In Section 3, we provided a brief overview of works that reconstructed speech from brain signals and explained the reason why employ the EEG-based AAD to support speech perception in a cocktail party environment. Followed with Sections 4 and 5, we introduced the conventional AAD algorithms and emerging works with deep learning, In Section 6, we focused on the application-oriented AAD works towards speech perception in cocktail parties. In Sections 7 and 8, we summarize the publicly available research resources for EEG-based AAD in cocktail party scenarios and discuss the challenges and future directions. We conclude this paper in Section 9.

2 Fundamental of the Speech Perception

Before delving into how a listener reacts to the cocktail party speech mixture, it will be helpful to review how the human brain completes the speech perception process. Speech perception occurs within a hierarchical processing system in the auditory system, involving several core brain regions [91]. A typical example is depicted in Figure 2. The speech is produced by the vocal folds and primarily

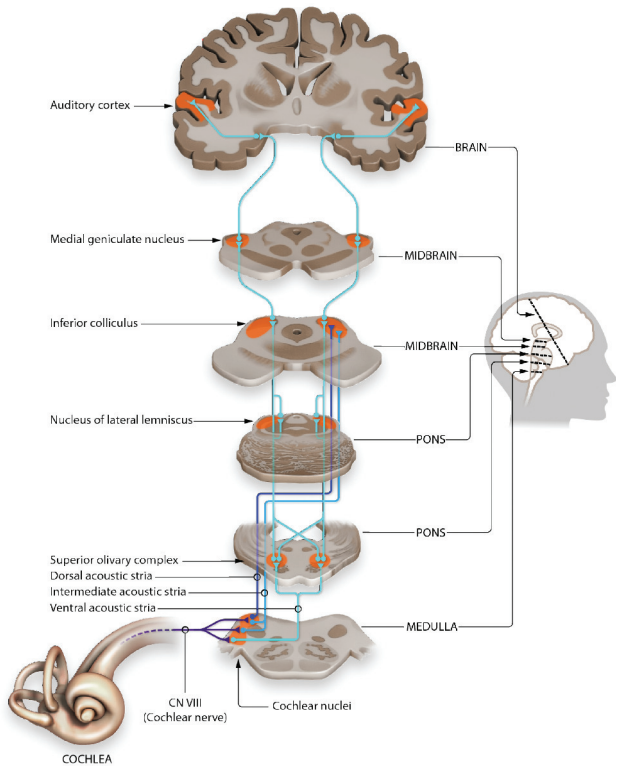

Figure 2: The auditory processing pathway. The speech perception begins with the cochlear nucleus and proceeds through a series of relay nucleus, including the superior olivary complex, the inferior colliculus, and the medial geniculate nucleus. Each of these nuclei decodes and integrates the incoming auditory information before forwarding it to the next stage of processing. Finally, the auditory cortex receives and analyzes the integrated signals, enabling the perception of speech. (Adopted from [91]).

received by the inner ear. The cochlea, as the main hearing organ in the inner ear, contains numerous nerve endings that convert sound vibrations into electrical impulses. These impulses, which correspond to different pitches or frequencies of sound, are then transmitted along the auditory nerve. These electrical impulses are further processed by various auditory nuclei, allowing for the estimation of physical characteristics such as spatial location cues by comparing signals from both ears. Finally, the auditory information reaches the auditory cortex, where the streams of nerve impulses are converted into meaningful sound, and multiple brain regions are engaged in the comprehension of speech.

Although the auditory system processes speech in a hierarchical manner, it cannot be solely regarded as a linear, feedforward process. Traditional

perception theories previously suggested that the brain processes stimuli in a bottom-up manner, constructing perceptions through the combination of sensory inputs [40]. However, recent theories propose that the brain is a dynamic system that interacts with sensory inputs [16]. This interaction is facilitated by the "top-down processing" mechanism, where perception is influenced by prior experiences and expectations. This top-down processing allows the brain to process sensory information more efficiently [89]. Furthermore, the interplay between bottom-up and top-down processing enables the brain to achieve precise perception, even in the presence of degraded sensory inputs [99].

In the field of speech perception, the brain is portrayed as a "prediction machine" where top-down expectations are constantly predicting bottom-up information [108]. Such a mechanism is especially noticeable in acoustically challenging scenarios. Despite potential degradation of the speech signal, such as background noise, overlapping speakers, or minor disruptions during communication, individuals with normal hearing can still, to a certain degree, follow the speaker. The predictive coding theory [13] offers a plausible explanation for this phenomenon, suggesting that top-down information generates prior expectations about speech content. This top-down information is constructed based on sensory inputs from various domains, such as speaker identity, speech knowledge, and language comprehension, and is further consolidated as cognitive factors. This also explains why, when people listen to strangers speak, their speech perception efficiency gradually improves over time, all due to the accumulation of the speaker's prior knowledge [91].

In the cocktail party problem, the top-down processing mechanism is also essential for speech perception. Previous studies have demonstrated that the brain responds to all stimuli, and top-down attention forces the neural activity to be selective in order to construct a representation only of the attended stream [3]. Consequently, decoding the target speaker in a noisy social environment necessitates extracting the relevant stream from the brain's signals, a task often accomplished using biological signals.

3 Speech Reconstruction from Brain Signal

Researchers have spent decades trying to figure out the neural representations of speech signals along the brain's auditory system. The stimulus reconstruction was proposed to interpret neural responses in the stimulus domain intuitively [4]. However, with the advent of the BCI concept, it is thought that reconstructing speech from the human auditory cortex is one of the ways for machines to establish direct communication with the brain. Considering the sampling rate gap between speech signals and brain signals, the reconstruction target is typically chosen as acoustic representations rather than the original waveforms.

The acoustic representations of the stimulus fall into two categories, i.e., discrete units and continuous speech. Common discrete units used include phonemes [68], phonetic categories [58], and words [67]. However, using discrete units eliminates prior knowledge provided by the top-down processing mechanism, such as paralinguistic information (e.g., speaker identity). Therefore, in this section, we concentrated on reconstructing the acoustic representations of continuous speech from continuous electronic brain signals. Previous studies primarily employed acoustic representations in either the time-frequency domain, such as the magnitude of the Short-Time Fourier Transform (STFT) spectrogram or Mel spectrogram, or the temporal domain, such as the envelope [44, 88]. Typically, brain signals are collected using multiple electrodes, making the stimulus reconstruction a multiple-input-multiple-output (MIMO) process when using time-frequency representations and a multiple-input-single-output (MISO) process when using temporal representations.

Regardless of representations, the concept of reconstruction can be divided into two categories: linear and nonlinear. The linear stimulus reconstruction method is first introduced in [79] and further adopted in various tasks, including the cocktail party problems. The fundamental idea of the linear stimulus reconstruction method is estimating a linear mapping between the acoustic representations and the population neural activity. With the spectrogram as an example, let us denote the ground-truth spectrogram as $S(t, f)$ and the reconstructed spectrogram as $\hat{S}(t, f)$. With the response at electrode n at time t as $R(t, n)$, the linear reconstruction is described as [79]:

$$\hat{S}(t, f) = \sum_{n} \sum_{\tau} g(\tau, f, n) R(t - \tau, n) \tag{1}$$

where $g(\tau, f, n)$ represents a spatio-temporal filter that maps $R(t, n)$ to $S(t, f)$. When the stimulus is an envelope, the filter becomes only temporal, and linear reconstruction is described as:

$$\hat{S}(t) = \sum_{n} \sum_{\tau} g(\tau, n) R(t - \tau, n) \tag{2}$$

The estimation of the filter $g(\cdot)$ is achieved by reducing the mean squared error (MSE) between the actual and reconstructed stimuli through the use of normalized reverse correlation.

By combining recent advances in deep learning, the nonlinear methods based on deep neural network (DNN) [4, 118] have significantly improved the reconstruction accuracy. In these methods, the stimulus reconstruction can be described by a composition of 2 networks with specific functions as:

$$\hat{S} = (\mathrm{A} \circ \mathrm{F})(R) \tag{3}$$

where $\mathrm{F}(\cdot)$ denotes the feature extraction network that reflects the neural responses R to high dimensional. $\mathrm{A}(\cdot)$ denotes the feature summation network

that nonlinearly regresses the high dimensional representations to the acoustic representations $\hat{S}$. On one hand, deep learning models have demonstrated their effectiveness in capturing statistical patterns of speech signals accurately. On the other hand, nonlinear regression methods have shown remarkable results in reconstructing the nonlinear encoding of speech features in neural data.

The works in the field of stimulus reconstruction also benefit speech perception in cocktail party problems. Research has shown that the human auditory system is capable of reconstructing the representation of the speaker being attended to and suppressing irrelevant speech, as if the person was listening to that speaker alone [62, 78, 88]. Besides, it has been verified that the selection of the acoustic representations of the stimulus [71] and the selection of frequency band of the brain signals [114] will significantly impact the estimation of the attentions in cocktail party problem.

In the prior studies, it was found that the attended speech envelope, that is a low-frequency component of the original speech, can be reconstructed from brain signals, such as ECoG or sEEG. The low-frequency signal can be used to identify the attended sound source, therefore, detecting the attended speaker. Unfortunately, ECoG or sEEG signals are collected from invasive devices, which is not practical for daily applications. The non-invasive EEG signals can be a convenient substitute [62, 70]. However, EEG has a lower signal-to-noise ratio, higher sensitivity to movement and artifacts, and lower bandwidth than ECoG or sEEG. Normally, most EEG studies focus primarily on the low-frequency range, including the δ ($<$ 4 Hz), θ (4-8 Hz), α (8-12 Hz), and β (13-30 Hz) bands that are commonly associated with speech production and perception in the human cortex [30]. Unfortunately, the gamma-band (around 70-150 Hz) within this range tends to be overlooked [103]. However, it is important to recognize that the γ-band has demonstrated a strong correlation with perception, cognitive function, and motor tasks [31].

To summarize, studies show that it is possible to reconstruct low-resolution speech stimuli from brain signals. This lays the foundation for auditory attention detection (AAD). By combining a speech separation module that separates multiple speakers from an input speech mixture [49], i.e. cocktail party, and an AAD module that detects and selects the attended speech or speaker, one may construct a neuro-steered hearing device.

4 Typical EEG-based Auditory Attention Detection

At a cocktail party, it's common to have multiple speakers talking at the same time. For ease of illustration, we only limit our discussion to two competing speakers, i.e. an attended and an unattended speaker. Most existing AAD models assume the availability of clean speech from the mixture of speakers

during run-time inference, so as to find the correlation between such clean speech and the EEG signals.

Let's denote a decision window of three time-aligned signals, i.e. the attended speech source, the unattended speech source, and the EEG signals as $\mathbf{s_0}$, $\mathbf{s_1}$, and $\mathbf{e}$ respectively. As will be discussed later, AAD aims to detect the attended speaker or locus index, which is denoted as $y \in \{0, 1\}$, representing one of the two speakers. As human attention may switch between the two speakers, a long speech-EEG signal can be segmented into a number of decision windows of length τ. The AAD function can be formulated as follows,

$$y = \mathbf{A}(\mathbf{s_0}; \mathbf{s_1}; \mathbf{e}) \tag{4}$$

where $\mathbf{A}(\cdot)$ is also called the window-wise AAD function in the rest of this paper. There are two typical ways to implement the AAD function, stimulus reconstruction or direct classification.

4.1 Stimulus Reconstruction

The stimulus reconstruction approach seeks to reconstruct the attended stimulus from the EEG signals and detect the attention in three steps.

1) Speech feature extraction

Since speech signals are sampled at a higher rate than EEG signals, it is essential to extract speech features that are synchronized with the EEG signals. Such monaural speech features can be represented by either one single signal (e.g., envelope [12, 88, 113]) or multiple signals (e.g., spectrogram). Let's denote the feature of $\mathbf{s_0}$ and $\mathbf{s_1}$ as $\mathbf{f_0}$ and $\mathbf{f_1}$, the feature extraction can be described by a function,

$$\mathbf{f_0} = \mathbf{F}(\mathbf{s_0}) \quad \mathbf{f_1} = \mathbf{F}(\mathbf{s_1}) \tag{5}$$

For instance, $\mathbf{F}(\cdot)$ could be a speech envelope or a spectrogram.

2) Stimulus reconstruction

The mapping between acoustic features of speech stimulus and observed EEG signals can be done in both ways [6, 74]. In this paper, we only discuss the reconstruction of acoustic features from EEG signals. Let's denote the output of the decoder as $\hat{\mathbf{f}}$, the mapping can be described as:

$$\hat{\mathbf{f}} = \mathbf{D}(\mathbf{e}) \tag{6}$$

The prediction function $\mathbf{D}(\cdot)$, also discussed in Section III, can be implemented by either linear regression or non-linear DNN.

3) Attention selection

With the reconstructed stimulus $\hat{\mathbf{f}}$ and the actual stimuli $\mathbf{f_0}$ and $\mathbf{f_1}$, one may easily detect the attended speech source by comparing through a similarity

function $\mathbf{C}(\cdot)$, e.g. a cosine similarity or Pearson correlation, we have $y_0 = \mathbf{C}(\hat{\mathbf{f}}, \mathbf{f_0})$ and $y_1 = \mathbf{C}(\hat{\mathbf{f}}, \mathbf{f_1})$.

$$y = \underset{m=\{0,1\}}{\operatorname{argmax}} y_m \tag{7}$$

4.2 Direct Classification

The direct classification approach [29] doesn't rely on the reconstructed stimulus. It employs a neural network $\mathbf{R}(\cdot)$ that takes the EEG signals and two speech features as input, and predicts the attended speaker through a regression function.

$$y = \mathbf{R}(\mathbf{f}_0, \mathbf{f}_1, \mathbf{e}) \tag{8}$$

$\mathbf{R}(\cdot)$ can be a neural network to perform the regression task. In [29], $\mathbf{R}(\cdot)$ is achieved with 2 convolutional layers and 4 fully connected layers, that are trained with the cross-entropy cost function.

The typical EEG-based AAD techniques are also reported in [29, 49]. They laid the foundation for the recent deep learning approaches, that are discussed next.

5 Deep Learning Approaches

With the advent of deep learning, several EEG-based AAD studies have reported superior performance to traditional methods. The success of deep learning approaches is built on the previous studies, namely stimuli reconstruction and direct classification, that can be summarized in three aspects.

1) Deep stimulus reconstruction

Deep learning models have shown superior performance for regression tasks in signal processing. The stimulus reconstruction task can be considered as an EEG-to-speech regression. It is generally believed that higher quality speech reconstruction leads to more accurate auditory attention detection for stimulus reconstruction approach to AAD. There have been recent studies exploring deep learning techniques for high-dimensional representations. For instance, the CNN-based vocoder [4], the dilated convolutional neural network [1, 2, 94], Long Short-Term Memory (LSTM) based [82] and, etc. With the improved modeling capability, these deep learning models improve the reconstruction quality with a short decision window, therefore, lower detection latency.

2) Deep EEG representation learning

Deep learning is known for its capability to learn representations that are highly effective for various downstream tasks, often outperforming traditional feature extraction or selection techniques. EEG signals pose significant challenges due to their high levels of noise and dimensionality, making traditional

feature representations less effective. In light of the successes of deep learning in signal processing and pattern classification, deep representation learning has emerged as a compelling alternative for EEG analysis.

One approach is the use of a CNN-based model as described in [82], which employs a data-driven approach to find the optimal representation. The other approach leverages prior knowledge from neuroscience to apply deep learning techniques and extract specific information from EEG signals. For example, a frequency-channel neural attention mechanism was introduced in [23] to dynamically assign differentiated weights to EEG signals based on their differing physiological origins. Additionally, the use of a Spiking Neural Network (SNN) has been explored in [18, 41] to learn the EEG representation from alpha power. The SNN is designed to imitate the neural computation and coding strategies in the brain, making it a promising approach for EEG representation.

3) Deep regression model

To associate EEG signals with speech stimuli, a regression model is often employed in either direct classification or stimulus reconstruction techniques. Notable examples of successful approaches include linear methods and non-linear neural networks. Deep neural networks represent the recent advances [19], where cross-model attention was used to dynamically adjust the weights of audio components based on the EEG attention vector, and show superior AAD performance.

However, it is important to note that EEG-based AAD is primarily studied in controlled laboratory settings with acoustic environments. To facilitate the application of EEG in real-world BCI systems, such as neuro-steered hearing devices, several implementation challenges must be addressed. In the following section, we will delve into these practical considerations and discuss their significance.

6 Towards Neuro-steered Hearing Devices

Figure 3 depicts a general diagram of a neuro-steered hearing device for speech perception. The microphone picks up a speech mixture, whereas the wearable EEG device records the corresponding EEG signals. The neuro-steered hearing device, guided by the EEG signals, seeks to extract the desired targeted speech from the mixture. In other words, the EEG-based AAD directs the speaker extraction mechanism to focus the attention on the target speaker. Therefore, a neuro-steered hearing device is also called neuro-steered speaker extraction. We next summarize the studies to overcome several implementation challenges.

- The clean target speech stimulus is unavailable during training and testing.

- The practical EEG signal acquisition on the move.
- The label of the attention is not available during training.

There have been studies in addressing the challenges, that also point to the emerging research directions as summarized next.

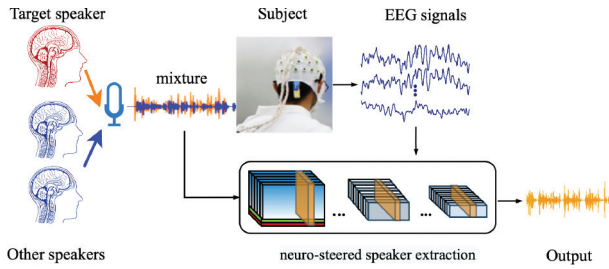

Figure 3: A neuro-steered hearing device for speech perception in a cocktail party environment.

6.1 Modeling without Clean Speech Stimuli

In a real-world acoustic environment, it is technically challenging to record the individual sound sources that make up a speech mixture. The individual sound sources are required as the reference during the training of an auditory attention detection model. Capturing each individual sound source during run-time inference is nearly impossible, which necessitates the development of neuro-steered speaker extraction solutions that can function without clean speech stimuli.

6.1.1 Training AAD Model with Separated Speech as Reference

A neuro-steered speaker extraction can be implemented by two parallel processes, speech separation, and auditory attention detection. Van Eyndhoven *et al.* [109] combined EEG-based auditory attention detection and non-negative blind source separation to effectively eliminate interfering sources, including the speaker not being attended to, from noisy multi-microphone recordings in a two-speaker acoustic environment. With the conventional envelope-reconstruction-based AAD and classical energy-based blind source separation, a system [11] was reported to show promising results in a cocktail environment. The viability of such an approach was further tested in [33], where the focus was on a binaural hearing aid in noisy environments with same-gender speakers positioned in different relative locations. Deep clustering [59] was used as the DNN-based speech separation algorithm instead of the previously utilized

training-free linear signal processing algorithm. The findings of the study indicated that AAD utilizing linear methods yielded comparable or superior performance compared to pure DNN-based methods. Multiple microphones were shown to improve speaker separation and the AAD performance over a single microphone. Despite the positive outcomes, one should note that both of these studies still necessitate the use of clean speech stimuli during training or calibration to develop the EEG decoder. This means that the need for clean speech stimuli has not been completely eliminated.

In real-world situations, it is not always necessary to separate all sources as a listener typically is interested in one of the speakers. Ceolini *et al.* [25] introduced the Brain-inspired Speech Separation (BISS) model which directly performs speech extraction, without the need of speech separation. In this study, a brain decoder is trained first to translate the brain signal into the speech envelope, which is then used as supplementary information along with short-time Fourier transform (STFT) features to train the extraction mask. Additionally, in [63], a Brain Enhanced Speech Denoiser (BESD) was proposed for end-to-end speech extraction from a mixture using Feature-wise Linear Modulation (FiLM) [92]. While the end-to-end training in BESD is simpler than the two-step approach in BISS, its performance is behind the state-of-the-art. Generally, the performance of deep learning-based speech extraction models is largely impacted by the availability of a large training dataset, which can be challenging to obtain using cocktail party datasets. In comparison, the extraction network in BISS was trained using a large artificial dataset, resulting in robust and good extraction performance.

In short, there have been studies to avoid the need of clean speech stimuli for auditory attention detection modeling. This can be achieved by working with a separately trained speaker extraction or speech separation model.

6.1.2 Training Spatial AAD Model without Speech Reference

Other than detecting the attended speaker, it is possible to detect the spatial location of the target speaker from EEG signals. It was found in neuroscience research that the location of auditory attention is reflected in brain activity [45, 114]. This has motivated the study of a particular type of EEG-based AAD to detect the spatial location of the target speaker, even in noisy or cluttered environments. This is also referred to as spatial AAD. There are two types of models, linear and nonlinear, in general. The spatial AAD takes a collection of EEG signals as input and predicts the contrastive spatial location of the attended speaker, e.g. left or right, front or rear. The training of such an AAD model doesn't rely on clean speech stimulus as the reference.

Bednar *et al.* [8] demonstrated that EEG responses to stimuli from different directions could be accurately classified using a support vector machine

(SVM) with a success rate significantly higher than the chance level of 25%. Furthermore, it was shown that EEG data could be utilized to track the path of an attended sound source with a linear reconstruction model [9]. Although a linear reconstruction model was employed in their study, its performance was inferior compared to envelope-based methods. Geirnaert *et al.* [51] implemented a data-driven linear filtering technique called filterbank common spatial pattern filters (FB-CSP) to achieve fast AAD, which outperformed the stimulus reconstruction approach in terms of accuracy on short signal segments. Further improvement was achieved by using a Riemannian geometry classifier instead of a traditional CSP filter [54].

Just like in many pattern classification tasks, the convolutional neural network (CNN) is an effective nonlinear model that detects the spatial focus of attention in multi-speaker scenarios [110]. The algorithm is effective in making accurate detection within 1-2 seconds. Feature representation is a crucial aspect of AAD, as raw EEG signals have low signal-to-noise ratios. To tackle this issue, Cai *et al.* [24] developed a method for spectro-spatial feature extraction in AAD using a CNN based on the alpha power's topographic specificity. This was followed by the development of the end-to-end spatiotemporal attention network (STAnet) [102]. STAnet integrates spatial and temporal attention mechanisms to capture both the modulation weights of EEG channels and the relevant temporal features of AAD. This spatiotemporal encoding method provides higher information density and outperforms traditional linear and non-linear methods on two widely used datasets. The use of deep learning in AAD has led to the development of more effective algorithms, and with continued advancements in feature extraction techniques, we can expect these algorithms to continue to evolve and make a significant impact in the field.

6.2 Simplifying EEG Acquisition

A sophisticated EEG cap is commonly required for EEG signal acquisition. For a practical neuro-steered hearing device, we call for a simplified EEG signal acquisition setup.

6.2.1 EEG Channel Selection

To simplify the standard EEG cap, it is desirable to remove some redundant EEG electrodes. The channel selection techniques are proven effective [81]. A low-density setup is expected to improve wearing comfort and reduce preparation time.

Mirkovic *et al.* [81] performed an iterative backward elimination algorithm to reduce electrodes from the initial electrode set and reported the first evidence that detection performance remains stable at a low number of EEG electrodes (from 96 channels to 25). Narayanan and Bertrand [83] developed a miniature

EEG device by using a greedy group-utility-based channel selection strategy and optimizing the channel combination through a mixed integer quadratic equation (MIQP) solver. They also examined the effect of reducing the inter-electrode distance and found that accuracy decreases significantly when the distance is less than 3 cm [85].

Unlike hard selection of EEG channels, some studies seek to adjust the weighting of EEG channels to derive more discriminative representations of AAD [22, 101], that is soft selection. This approach leverages the fact that some channels provide more insight into the brain's decision-making process in AAD, while others may provide less information. By assigning different weights to different channels, soft selection takes full advantage of the information provided by all channels, resulting in a more complete picture of AAD. In comparison to the hard selection, the soft selection is better suited to handle the variability and complexity of EEG signals, which can often be difficult to capture using a fixed set of channels. By taking a more flexible approach, the soft selection is able to account for the variability of EEG signals and provide a more accurate representation of AAD.

6.2.2 Ear-EEG

In practical AAD tasks, the EEG signals are acquired from the subjects in a real-world environment as opposed to a controlled setup in the lab. Unfortunately, conventional scalp EEG data collection is typically done in the lab, which is cumbersome and unsuitable for mobile applications. To address this, ear-EEG has been developed as an alternative to traditional scalp EEG. It provides less coverage of the brain but has the benefit of being more convenient and portable.

As shown in Figure 4 (a), in-ear EEG places multiple electrodes in the external auditory canal and over the outer ear through individualized earplugs [73]. Several studies have demonstrated that relevant neural signals can be successfully extracted from in-ear EEG recordings for AAD purposes [15, 65]. Despite having weaker amplitude compared to scalp EEG recordings [42, 43], in-ear EEG provides a convenient, portable, and virtually unnoticeable solution.

The around-the-ear EEG approach utilizes electrodes that are positioned close to the ear in a circular configuration around the outer ear, as shown in Figure 4 (b). Debener *et al.* [36] introduced the first flexible, printed Ag/AgCl electrode system with 10 electrodes arranged in a c-shape to fit comfortably on the ear. It's called cEEGrid and offers a promising solution for neuro-steered hearing aids. A series of validation studies demonstrated that the cEEGrid could achieve reliable ear-EEG recordings [14, 66, 77]. Denk et al. conducted a comparison between the signal properties of around-the-ear and in-ear EEG electrodes [38]. They found that around-the-ear electrodes had

(a) (b)

Figure 4: Illustration of the ear-EEG design and layout (a) Ear-EEG in the ear (adopted from [15]) (b) Ear-EEG around the ear. (Adopted from [36]).

several advantages over in-ear electrodes, including larger amplitude, improved channel independence, and a higher signal-to-noise ratio (SNR). These benefits were attributed to the greater inter-electrode distance in around-the-ear EEG recordings.

Inspired by these findings, several research explored whether ear-EEG recorded by cEEGrids can be used for detecting auditory attention. The study by Mirkovic *et al.* [80] was the first to show that the cEEGrid ear-EEG method can detect the attended speaker with an average accuracy of 69.3%, which is above the chance level. On the other hand, the 84-channel cap-EEG resulted in an accuracy of 84.8%. This difference in accuracy can be attributed to the signal loss from the scalp to the ear. Specifically, Meiser and Bleichner [76] found that cEEGrid ear-EEG recordings showed a reduction in signal loss of 21% to 44% for four different auditory ERPs (N100, MMN, P300, and N400), when compared to 96-channel cap-EEG. Despite the lower AAD accuracy, the cEEGrid method has a practical advantage as it is portable, operated using a smartphone, and nearly invisible [64]. It is worth noting that Holtze *et al.* [61] modified the cEEGrid ear-EEG method inčitemirkovic2016target with individually chosen hyperparameters and significantly improved AAD performance. This also suggests that cEEGrid has the potential to be considered as a suitable EEG acquisition tool for use in neural-guided hearing aids, and thus, deserves further investigation to improve its performance.

6.3 Unsupervised Learning

Training an AAD model, one may expect that the label of the attended speaker is known. However, such a label collection procedure is labor-intensive and less practical in hearing devices. Therefore, unsupervised learning could be

an alternative [52, 53]. The first AAD work based on stimuli reconstruction approach with unsupervised learning is proposed in [53]. It assumes that only the two envelopes of the competing speakers and EEG data are presented during the training phase.

As it has been verified that the brain encodes attended and unattended speakers differently, it is possible to identify which envelope is the attended one and which one is unattended by using a decoder. This can be done by iteratively replacing the ground truth attention labels utilized in supervised training with the predicted labels obtained from the testing phase. This creates a self-reinforcing effect, where each iteration improves the decoder's performance, even in the presence of labeling errors. This idea has been expanded upon in a study by Geirnaert et al. [52], who developed a time-adaptive, unsupervised stimulus reconstruction method that operates online. The method continually adjusts and improves itself as new EEG and audio data streams in, through the use of sliding window training or recursive training. Both of these methods perform better than traditional time-invariant supervised decoders.

6.4 AAD for the Hearing-Impaired

While EEG-based AAD is mostly studied for normal hearing (NH) subjects, neuro-steered hearing devices could assist hearing-impaired (HI) subjects as well. In most of the publicly available EEG-based AAD datasets, the listening subjects are mostly young and normal-hearing people, that don't represent the demography of the hearing impaired. Globally, 34 million children are deaf or have hearing loss, and approximately 30% of people over the age of 60 have hearing loss [26]. Caution should be taken before applying previous results or technologies to detect auditory attention in subjects with hearing impairments.

Hearing loss in children can be present at birth (congenital) or develop later in childhood (acquired). As for elderly people, several studies have shown that, in addition to peripheral hearing loss, speech perception is also affected by changes in brain structure as well as changes in brain function [57, 106, 115]. Given that aging and hearing loss are major causes of neuromodulation decline in the listening brain [57, 107], AAD research needs to take these effects into account. Nogueira et al [86] first compared the EEG-based AAD performance of NH and HI listeners. 12 NH listeners (age: 26 $\pm$ 4.4 years) and 12 bilateral implanted cochlear implant users (age: 60 $\pm$ 11.0 years) were involved in this study. Results demonstrated that in principle it is possible to detect selective attention in individuals with HI with an accuracy of up to 70%, while an average accuracy of NH listeners is higher than 80%. This is also supported by that the HI listeners rated the competing speech task to be more difficult [46]. Meanwhile, some studies have also verified the feasibility of

detecting selective attention through the ear-EEG signals of HI subjects [48, 55, 87].

6.5 Implementation of Practical Hearing Devices

To bring the above AAD research from the laboratory to real life, there are many other challenges that we need to overcome.

- It is known that movements and distractions are reflected in brain signals associated with auditory attention during daily life. However, most AAD research has so far been performed in controlled laboratory settings. This potentially limits the generalizability of existing studies to complex acoustic environments outside the laboratory.

- A practical system calls for real-time detection of auditory attention. Usually, a complex model leads to high accuracy when operating at high temporal resolutions. However, the computing cost and the limited data resource need to be taken into consideration. We need to find a tradeoff between the model's complexity and accuracy. Furthermore, real-time implementation is a causal system that can only use historical data. That is different from the offline system.

- The coupling between the AAD system and the hearing device is also a challenge. A fully mobile EEG recording system is needed. Moreover, a compact design that includes brain signal acquisition, processing, speech signal acquisition, and processing units, and their communications. Furthermore, it usually introduces communication delays between hardware components.

However, this study is still in its infancy. In [60], cEEGrids were placed near the left and right ears of the participants to record ear-EEG signals related to speech perception. For the first time, this was done for a continuous period of six hours as the participants carried out various activities such as working on a computer, conversing with coworkers, and taking lunch breaks, while also performing auditory oddball tasks in and out of the laboratory. The results indicated that the participants were able to differentiate between target and non-target sounds even while engaged in their daily activities. Additionally, it was found that the participants had higher ERP amplitudes in response to target tones as compared to standard tones. These findings suggest that it is possible to study auditory attention outside of traditional laboratory settings.

Inspired by this, a hearing aid-EEG research platform has been developed in [34, 35]. As shown in Figure 5, the Portable Hearing Laboratory (PHL) is a comprehensive hearing aid research system that can be used to present auditory stimuli to subjects and perform low-latency audio signal processing.

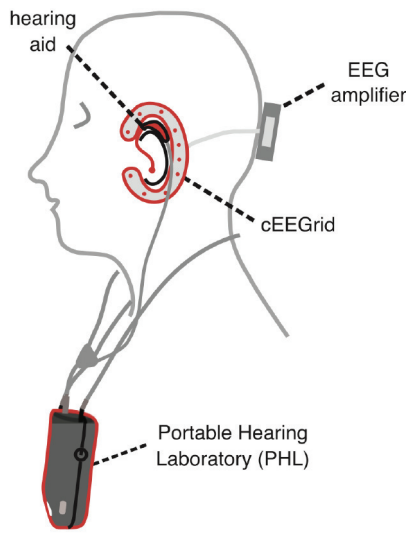

Figure 5: A hearing aid with ear-EEG recordings platform (Adopted from [35].)

As for the EEG system, it consists of cEEGrid, a mobile EEG amplifier, and a smartphone. Therefore, both audio and EEG data can be acquired and synchronized using this hearing aid-EEG research device. While the PHL and cEEGrid are not entirely suitable for daily use, this setup provides a potential platform for exploring closed-loop EEG & audio applications in a research context.

7 Datasets for Auditory Attention Detection Study

In recent years, there are few EEG datasets suitable for AAD research. Typically these datasets were collected from a dozen of young normal-hearing subjects. To reduce the cognitive load, it is best that we recruit speakers to listen to their native languages. Since the data were recorded in different countries, the languages vary from dataset to dataset. The characteristics of all publicly available data sets are summarized in Table 1.

The collection of AAD datasets typically follows a similar procedure. Take ESAA [20, 21] as an example, the participants were instructed to focus on one speaker while disregarding the other in a scenario with two overlapping speakers. The speech material consisted of various Chinese narratives narrated by two native speakers and was normalized to have the same root mean squared intensity, making the stimuli appear equally loud. The stimuli were processed using a head-related transfer function (HRTF) to simulate speech sources

Table 1: The characteristics of different AAD datasets. NH = Normal-hearing, HI = Hearing-impaired.

Dataset	NH subjects	HI subjects	Language	# cap-EEG channels	Duration per subject (min)
Das-2015 [32]	16	0	Flemish	64	48
Fuglsang-2018 [47]	18	0	Danish	64	50
Fuglsang-2020 [46]	22	22	Danish	64[a]	40
ESAA [21]	20	0	Chinese	64	38
Neural Tracking to go [100]	20	0	German	24[b]	30

Note: [a]In-ear EEG was also recorded for 19 of the 44 subjects. [b]A fully mobile EEG Device.

located at 90-degree intervals to the left and right of the subjects. EEG data was acquired using a BrainAmp system operating at a sampling rate of 8,192 Hz, with a 64-channel recording setup. To ensure that the participants were attentive during the experiment, participants were asked to complete a multiple-choice questionnaire following each trial to assess their comprehension of two separate narratives. To avoid fatigue or loss of focus, participants were given short breaks after each trial and longer breaks after 8 consecutive trials. To control for potential biases, the position of the target streams and the gender of the speakers were randomized for each participant throughout the course of the experiments.

Overall, the size of AAD datasets is highly limited, especially in the context of deep learning. This calls for a great effort in data collection. Furthermore, there is a rising interest in augmenting speech processing tasks with additional biological signals [97], such as surface electromyography (sEMG) and Electrooculography (EOG) signals. The integration of these multimodal physiological signals holds promise for AAD-enabled applications. Such studies rely on multimodal physiological signals to unlock the AAD potential.

8 Challenges and Directions

EEG-based speech perception has achieved promising success in solving the cocktail party problem, but there are challenges associated with its adoption for neuro-steered hearing devices that merit further discussion.

8.1 Process of the Noisy EEG Signals

Several traditional signal processing techniques have been studied in the area of EEG-based speech perception, as demonstrated in recent review articles [29, 49]. Despite these efforts, the EEG signal still exhibits a poor signal-to-noise ratio, resulting in limited success in achieving optimal AAD performance.

Deep learning techniques that process raw input data are referred to as representation learning methods. The objective of representation learning is to extract the most significant features from raw data, leading to improved pattern recognition performance [10]. Unfortunately, most previous AAD architectures have not benefited from representation learning. In recent years, some researchers have suggested the use of advanced deep learning methods [23, 102], which have the potential to extract discriminative features that can improve AAD performance. Therefore, it is worth further investigating deep learning frameworks that extract representations directly needed for classification or detection from raw EEG signals. Specifically, to extract relevant attention features from raw EEG, information about the EEG signals needed to be further exploited, e.g., the characteristics in the time, frequency, and spatial domains. However, because most deep learning models benefit from large model sizes, how to adapt them to AAD's small dataset remains a challenge. Transfer learning has gained popularity in EEG signal processing as a potential solution to overcome the limitations of small datasets [111]. This technique involves utilizing pre-trained models that have been extensively trained on larger EEG datasets or similar tasks. The pre-trained models are then fine-tuned using smaller, task-specific EEG datasets. By leveraging the knowledge acquired during pre-training, transfer learning allows for rapid adaptation towards specific EEG-based AAD tasks [117].

8.2 Generalization of AAD models

There are two main aspects of generalization. One is the generalization across subjects, another is the generalization across scenarios.

The variability of brain signals in individuals presents a challenge for EEG-based BCI systems, particularly in subject-independent conditions. Brain signals of each individual can change over time due to differences in their physiological and psychological traits [69]. Additionally, the unique spatial origin, amplitude, and variability of brain signals can make it difficult to detect auditory attention tasks in a subject-independent manner [95]. In general, traditional auditory attention detection methods in EEG-based BCI systems work well in subject-dependent conditions but struggle in subject-independent conditions. This may be due to the fact that brain signals from different individuals are highly variable, discriminative, and carry specific meaning in auditory attention detection tasks. To compensate for these variations, BCI systems often require a calibration process, which adds an extra burden for the user and hinders the practical use of BCIs.

In addition to generalization across subjects, generalization across scenarios is also a challenge for EEG-based BCI systems. Most AAD studies are conducted in controlled laboratory settings, which limits the generalization of findings to complex acoustic environments in real-life scenarios. In real-life

situations, subjects often handle multiple tasks, which are reflected in their brain activities. Selective listening is just one of these tasks, and detecting auditory attention from the mixture of brain activities in EEG signals is an area that requires further study.

8.3 Complexity, Cost, and Tracking Latency

In most auditory attention studies, we make a decision based on a single window of the signal. However, selective listening in the human brain is a continuous process. It is likely that human cognitive resources, loaded by the cognitive process of auditory attention to the speech sources, are affected by its previous cognitive states, as well as other factors such as distracting auditory events, moving auditory events, or other cognitive and motor activities. In short, tracking auditory attention is of practical need and yet an unexplored challenging research problem.

In general, deep neural networks have a high demand for energy consumption, data requirements, and computational power. However, these demands are particularly pronounced in BCI applications, including neuro-steered hearing devices, due to the limited data size, energy supply, and the need for real-time response. To address these challenges, various hardware accelerators have been developed to manage the high computational demands of deep learning models. Despite these advances, there remains a need for a low-cost, energy-efficient AAD algorithm that can be implemented on a single chip.

The human brain is a sophisticated network of neurons and synapses that transmit information through electrical impulses referred to as spikes. This remarkable processing capability has led to the evolution of spiking neural networks (SNNs) as a potentially valuable computing framework. SNNs operate by allowing neurons to communicate with each other through spikes with adjustable weight values that are transmitted via synapses connecting the neurons [72]. Research has demonstrated that the low computational cost of SNNs makes them well-suited for deployment on low-power hardware [104].

Considering that the AAD model is built to process brain signals, a brain-like model should be a natural choice. The utilization of SNNs in AAD offers several advantages over traditional deep learning models. Unlike deep learning models, SNNs are capable of processing data in real-time and can effectively handle noisy and unstructured data. Additionally, SNNs consume significantly less power compared to deep learning models, making them ideal for practical deployment. The ability of SNNs to simulate the dynamic nature of biological neurons and model the temporal relationships between spikes is also beneficial for AAD applications.

In conclusion, SNNs represent a promising computing paradigm that offers several advantages over traditional deep learning models. The low computational cost, real-time processing capabilities, and ability to handle noisy and

unstructured data make SNNs a suitable choice for AAD applications. Further research in the field of SNN-based AAD is encouraged, as it has the potential to lead to exciting new developments in this field.

8.4 Connection between Speech Separation and AAD

As the two sub-fields of speech perception in cocktail party environments, the development speed of speech separation and AAD are imbalanced, especially in the deep learning era. For speech separation, it is easy to generate the mixture from the public speech corpus, leading to a low cost of data acquisition. Therefore, current speech separation models employ deep learning to improve representative learning and dependency modeling and achieve amazing success.

However, for the AAD, these sophisticated models are not so easy to exploit. Given the collection of EEG data is labor-intensive, the small size of the AAD dataset makes the training of the deep learning models easy to overfit. Besides, although assistive speech perception can be achieved by the pipeline of speech separation and AAD, it is not clear whether the two local optimal achieve the global optimal. In other words, the impact of the artifacts introduced by speech separation on the AAD is not clear. The end-to-end training could be a potential solution, however, as indicated in [25], the performance is still limited by the data size. Although the separation data are easy to make, it is almost impossible to have the equivalent EEG data. Besides, the separation module usually introduces the permutation problem, which means the sequence of the output separated speech is usually randomly which may affect the performance of the AAD training. Although permutation invariant training (PIT) [119] can be adopted in the training stage, it cannot be employed in the online separation phase given the latency issue.

Moreover, there is no consensus on how to evaluate these pipeline systems. As the performance of speech separation is measured by SI-SNR whereas the AAD is evaluated with accuracy. In some cases, System A might perform better than System B on separation performance but worse on AAD accuracy, the comparison becomes confusing. In addition, as the AAD accuracy is closely related to the length of the decision window, a longer decision window indicates higher accuracy, but is less sensitive to attention shifting. Thus, from the perspective of speech perception, using the accuracy of AAD as a direct evaluation metric may also not be appropriate. Besides, the gain control of speeches for continuous decoding in a cocktail party environment generates another problem. As sudden switching of speakers (of which many by mistake) cause perceptually unpleasant spurious. Although an interpretable performance metric for AAD algorithms has been developed with adaptive gain controls in [50] using a Markov chain model, which assumes independence between consecutive decisions. However, in real-world applications, data is often segmented with overlapping in order to reduce processing latency. As

a result, the connection between the deep-learning-based speech separation model and AAD in the design of neuro-steered hearing devices remains an open question.

9 Conclusion

This paper provides a comprehensive overview of EEG-based AAD for speech perception in noisy environments, such as cocktail party scenarios. It covers the essential concepts and the latest developments in the field up to 2023. The paper explores the underlying mechanisms of speech perception and the ways to build a machine that mimics the human brain to solve the cocktail problem. Additionally, the paper provides an overview of the current deep learning approaches in the field, discussing their potential and limitations. Furthermore, it points out the gap between EEG-based speech perception research and neuro-steered hearing device, and provides a list of resources available to advance the research. Overall, this article is a valuable resource for anyone interested in comprehending EEG-based auditory attention detection and developing novel deep learning techniques.

Financial Support

This work is supported by A∗STAR under its RIE 2020 Advanced Manufacturing and Engineering Programmatic Grant (Grant No. A1687b0033), and by Internal Project of Shenzhen Research Institute of Big Data (Grant No. T00120220002); the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen (Grant No. B10120210117-KP02), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany).

Biographies

Siqi Cai received her Ph.D. degree from the Department of Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, China, in 2020. She is now a Research Fellow at the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Her research interests include brain-computer interface, and biosignal processing. She has served as the workshop chair of the 47th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2022.

Hongxu Zhu obtained his bachelor's degree in information engineering from Xi'an Jiaotong University, China. He further obtained his M.S. with distinction in electronic information engineering and Ph.D. in electrical engineering from City University of Hong Kong in 2016 and 2021, respectively. Currently, he is a research fellow with the Human Language Technology Lab in the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He is also a key member of the IEEE Standards Working Group P2668 and P1451.5. His research interests broadly lie in brain-informed speaker separation, machine learning, and internet of things.

Tanja Schultz received her doctoral and diploma degree in Informatics from University of Karlsruhe, Germany, in 2000 and 1995. She joined Carnegie Mellon University, Pittsburgh, PA in 2000 and is an adjunct Research Professor at the Language Technologies Institute. From 2007 to 2015 she was a Full Professor in Informatics at the Karlsruhe Institute of Technology (KIT) in Germany before she became a Professor for Cognitive Systems at the University of Bremen, Germany in April 2015. Since 2007, she directs the Cognitive Systems Lab, where her research activities focus on the processing, recognition, and interpretation of biosignals for human-centered technologies and applications. She is an ISCA Fellow and member of the European Academy of Sciences and Arts.

Haizhou Li received the B.Sc., M.Sc., and Ph.D degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. Dr Li is currently a Professor at Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China, and Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. He was the General Chair of ACL 2012, INTERSPEECH 2014, ASRU 2019, and ICASSP 2022. Dr Li is a Fellow of the IEEE and the ISCA. He was a recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was the President of Asia Pacific Signal and Information Processing Association (2015–2016). He is currently a Vice President of IEEE Signal Processing Society (2024–2026).

References

[1] B. Accou, M. J. Monesi, T. Francart, *et al.*, "Predicting speech intelligibility from EEG in a non-linear classification paradigm," *Journal of Neural Engineering*, 18(6), 2021, 066008.

[2] B. Accou, M. J. Monesi, J. Montoya, T. Francart, *et al.*, "Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network," in *2020 28th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, 1175–9.

[3] J. Ahveninen, M. Hämäläinen, I. P. Jääskeläinen, S. P. Ahlfors, S. Huang, F.-H. Lin, T. Raij, M. Sams, C. E. Vasios, and J. W. Belliveau, "Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise," *Proceedings of the National Academy of Sciences*, 108(10), 2011, 4182–7.

[4] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific reports*, 9(1), 2019, 1–12.

[5] S. Akram, A. Presacco, J. Z. Simon, S. A. Shamma, and B. Babadi, "Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling," *NeuroImage*, 124, 2016, 906–17.

[6] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A Tutorial on Auditory Attention Identification Methods," *Frontiers in Neuroscience*, 13, 2019, DOI: 10.3389/fnins.2019.00153, https://www.frontiersin.org/articles/10.3389/fnins.2019.00153.

[7] W. M. H. Bakay, L. A. Anderson, J. A. Garcia-Lazaro, D. McAlpine, and R. Schaette, "Hidden hearing loss selectively impairs neural adaptation to loud sound environments," *Nature Communications*, 9(1), 2018, 1–11.

[8] A. Bednar, F. M. Boland, and E. C. Lalor, "Different spatio-temporal electroencephalography features drive the successful decoding of binaural and monaural cues for sound localization," *European Journal of Neuroscience*, 45(5), 2017, 679–89.

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

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 2013, 1798–828.

[11] A. Bertrand and M. Moonen, "Energy-based multi-speaker voice activity detection with an ad hoc microphone array," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, 85–8.

[12] 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*, 25(5), 2016, 402–12.

[13] H. Blank and M. H. Davis, "Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception," *PLoS biology*, 14(11), 2016, e1002577.

[14] M. G. Bleichner and S. Debener, "Concealed, unobtrusive ear-centered EEG acquisition: cEEGrids for transparent EEG," *Frontiers in Human Neuroscience*, 11, 2017, 163.

[15] M. G. Bleichner, M. Lundbeck, M. Selisky, F. Minow, M. Jäger, R. Emkes, S. Debener, and M. De Vos, "Exploring miniaturized EEG electrodes for brain-computer interfaces. An EEG you do not see?" *Physiological Reports*, 3(4), 2015, e12362.

[16] J. N. de Boer, M. M. Linszen, J. de Vries, M. J. Schutte, M. J. Begemann, S. M. Heringa, M. M. Bohlken, K. Hugdahl, A. Aleman, F. N. Wijnen, *et al.*, "Auditory hallucinations, top-down processing and language perception: a general population study," *Psychological medicine*, 49(16), 2019, 2772–80.

[17] A. W. Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, 77(5), 2015, 1465–87.

[18] S. Cai, P. Li, E. Su, Q. Liu, and L. Xie, "A Neural-Inspired Architecture for EEG-Based Auditory Attention Detection," *IEEE Transactions on Human-Machine Systems*, 52(4), 2022, 668–76, DOI: 10.1109/THMS.2022.3176212.

[19] S. Cai, P. Li, E. Su, and L. Xie, "Auditory Attention Detection via Cross-Modal Attention," *Frontiers in Neuroscience*, 15, 2021.

[20] S. Cai, E. Su, P. Li, J. Li, L. Xie, and H. Li, "ESAA: An EEG-Speech Auditory Attention Detection Database," in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, IEEE, 2022, 1–6, DOI: 10.1109/O-COCOSDA202257103.2022.9997944.

[21] S. Cai, E. Su, P. Li, J. Li, L. Xie, and H. Li, *ESAA: an EEG-Speech auditory attention detection database*, version 1.0, Zenodo, September 2022, DOI: 10.5281/zenodo.7078451, https://doi.org/10.5281/zenodo.7078451.

[22] S. Cai, E. Su, L. Xie, and H. Li, "EEG-Based Auditory Attention Detection via Frequency and Channel Neural Attention," *IEEE Transactions on Human-Machine Systems*, 52(2), 2021, 256–66.

[23] S. Cai, E. Su, L. Xie, and H. Li, "EEG-Based Auditory Attention Detection via Frequency and Channel Neural Attention," *IEEE Transactions on Human-Machine Systems*, 52(2), 2022, 256–66.

[24] S. Cai, P. Sun, T. Schultz, and H. Li, "Low-latency auditory spatial attention detection based on spectro-spatial features from EEG," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, 5812–5.

[25] E. Ceolini, J. Hjortkjær, D. D. 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*, 223, 2020, 117282.

[26] S. Chadha, K. Kamenov, and A. Cieza, "The world report on hearing, 2021," *Bulletin of the World Health Organization*, 99(4), 2021, 242.

[27] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, 25(5), 1953, 975–9.

[28] 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*, 172, 2018, 206–16.

[29] 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*, 9(1), 2019, 1–10.

[30] N. E. Crone, D. Boatman, B. Gordon, and L. Hao, "Induced electrocorticographic gamma activity during auditory perception," *Clinical Neurophysiology*, 112(4), 2001, 565–82.

[31] N. E. Crone, A. Sinai, and A. Korzeniewska, "High-frequency gamma oscillations and human brain mapping with electrocorticography," *Progress in Brain Research*, 159, 2006, 275–95.

[32] N. Das, T. Francart, and A. Bertrand, *Auditory Attention Detection Dataset KULeuven*, Zenodo, August 2020, DOI: 10.5281/zenodo.3997352, https://doi.org/10.5281/zenodo.3997352.

[33] N. Das, J. Zegers, T. Francart, A. Bertrand, *et al.*, "Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding," *Journal of Neural Engineering*, 17(4), 2020, 046039.

[34] S. Dasenbrock, S. Blum, S. Debener, V. Hohmann, and H. Kayser, "A step towards neuro-steered hearing aids: Integrated portable setup for time-synchronized acoustic stimuli presentation and EEG recording," *Current Directions in Biomedical Engineering*, 7(2), 2021, 855–8.

[35] S. Dasenbrock, S. Blum, P. Maanen, S. Debener, V. Hohmann, and H. Kayser, "Synchronization of ear-EEG and audio streams in a portable research hearing device," *Frontiers in Neuroscience*, 16, 2022.

[36] S. Debener, R. Emkes, M. De Vos, and M. Bleichner, "Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear," *Scientific Reports*, 5(1), 2015, 1–11.

[37] Y. Deng, I. Choi, and B. Shinn-Cunningham, "Topographic specificity of alpha power during auditory spatial attention," *NeuroImage*, 207, 2020, 116360.

[38] F. Denk, M. Grzybowski, S. M. Ernst, B. Kollmeier, S. Debener, and M. G. Bleichner, "Event-related potentials measured from in and around the ear electrodes integrated in a live hearing device for monitoring sound perception," *Trends in Hearing*, 22, 2018, 2331216518788219.

[39] 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*, 109(29), 2012, 11854–9.

[40] A. K. Engel, P. Fries, and W. Singer, "Dynamic predictions: oscillations and synchrony in top–down processing," *Nature Reviews Neuroscience*, 2(10), 2001, 704–16.

[41] F. Faghihi, S. Cai, and A. A. Moustafa, "A neuroscience-inspired spiking neural network for EEG-based auditory spatial attention detection," *Neural Networks*, 152, 2022, 555–65.

[42] L. Fiedler, J. Obleser, T. Lunner, and C. Graversen, "Ear-EEG allows extraction of neural responses in challenging listening scenarios—a future technology for hearing aids?" In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2016, 5697–700.

[43] L. Fiedler, M. Wöstmann, C. Graversen, A. Brandmeyer, T. Lunner, and J. Obleser, "Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech," *Journal of Neural Engineering*, 14(3), 2017, 036020.

[44] T. Francart, N. Das, S. Van Eyndhoven, W. Biesmans, and A. Bertrand, "Neuro-steered noise suppression for auditory prostheses," *The Journal of the Acoustical Society of America*, 139(4), 2016, 2044–4.

[45] J. N. Frey, N. Mainy, J.-P. Lachaux, N. Müller, O. Bertrand, and N. Weisz, "Selective modulation of auditory cortical alpha activity in an audiovisual spatial attention task," *Journal of Neuroscience*, 34(19), 2014, 6634–9.

[46] 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*, 40(12), 2020, 2562–72.

[47] S. A. Fuglsang, D. D. Wong, and J. Hjortkjær, *EEG and Audio Dataset for Auditory Attention Decoding*, version 1, Zenodo, March 2018, DOI: 10.5281/zenodo.1199011, https://doi.org/10.5281/zenodo.1199011.

[48] M. Garrett, S. Debener, and S. Verhulst, "Acquisition of subcortical auditory potentials with around-the-ear cEEGrid technology in normal and hearing impaired listeners," *Frontiers in Neuroscience*, 13, 2019, 730.

[49] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices," *IEEE Signal Processing Magazine*, 38(4), 2021, 89–102.

[50] S. Geirnaert, T. Francart, and A. Bertrand, "An Interpretable Performance Metric for Auditory Attention Decoding Algorithms in a Context of Neuro-Steered Gain Control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2019.

[51] 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*, 68(5), 2020, 1557–68.

[52] S. Geirnaert, T. Francart, and A. Bertrand, "Time-adaptive Unsupervised Auditory Attention Decoding Using EEG-based Stimulus Reconstruction," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[53] S. Geirnaert, T. Francart, and A. Bertrand, "Unsupervised Self-Adaptive Auditory Attention Decoding," *IEEE Journal of Biomedical and Health Informatics*, 25(10), 2021, 3955–66.

[54] Geirnaert, Simon and Francart, Tom and Bertrand, Alexander, "Riemannian geometry-based decoding of the directional focus of auditory attention using EEG," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, 1115–9.

[55] M. Geravanchizadeh and S. Zakeri, "Ear-EEG-based binaural speech enhancement (ee-BSE) using auditory attention detection and audiometric characteristics of hearing-impaired subjects," *Journal of Neural Engineering*, 18(4), 2021, 0460d6.

[56] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, 17(9), 2005, 1875–902.

[57] M. J. Henry, B. Herrmann, D. Kunke, and J. Obleser, "Aging affects the balance of neural entrainment and top-down neural modulation in the listening brain," *Nature Communications*, 8(1), 2017, 1–11.

[58] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: a focused review," *Frontiers in Neuroscience*, 10, 2016, 429.

[59] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, 31–5.

[60] D. Hölle, J. Meekes, and M. G. Bleichner, "Mobile ear-EEG to study auditory attention in everyday life," *Behavior research methods*, 53(5), 2021, 2025–36.

[61] B. Holtze, M. Rosenkranz, M. Jaeger, S. Debener, and B. Mirkovic, "Ear-EEG Measures of Auditory Attention to Continuous Speech," *Frontiers in Neuroscience*, 2022, 539.

[62] C. Horton, R. Srinivasan, and M. D'Zmura, "Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party'," *Journal of Neural Engineering*, 11(4), 2014, 046015.

[63] M. Hosseini, L. Celotti, and É. Plourde, "Speaker-Independent Brain Enhanced Speech Denoising," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, 1310–4, DOI: 10.1109/ICASSP39728.2021.9414969.

[64] M. Jaeger, B. Mirkovic, M. G. Bleichner, and S. Debener, "Decoding the attended speaker from EEG using adaptive evaluation intervals captures fluctuations in attentional listening," *Frontiers in Neuroscience*, 14, 2020, 603.

[65] D.-H. Jeong and J. Jeong, "In-ear EEG based attention state classification using echo state network," *Brain Sciences*, 10(6), 2020, 321.

[66] S. L. Kappel, S. Makeig, and P. Kidmose, "Ear-EEG forward models: improved head-models for ear-EEG," *Frontiers in Neuroscience*, 13, 2019, 943.

[67] S. Kellis, K. Miller, K. Thomson, R. Brown, P. House, and B. Greger, "Decoding spoken words using local field potentials recorded from the cortical surface," *Journal of neural engineering*, 7(5), 2010, 056007.

[68] B. Khalighinejad, G. C. da Silva, and N. Mesgarani, "Dynamic encoding of acoustic features in neural responses to continuous speech," *Journal of Neuroscience*, 37(8), 2017, 2176–85.

[69] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain–computer interfaces based on deep convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 31(10), 2019, 3839–52.

[70] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European Journal of Neuroscience*, 31(1), 2010, 189–93.

[71] G. M. D. Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing," *Current Biology*, 25(19), 2015, 2457–65, DOI: https://doi.org/10.1016/j.cub.2015.08.030, https://www.sciencedirect.com/science/article/pii/S0960982215010015.

[72] J. L. Lobo, J. Del Ser, A. Bifet, and N. Kasabov, "Spiking neural networks and online learning: An overview and perspectives," *Neural Networks*, 121, 2020, 88–100.

[73] D. Looney, C. Park, P. Kidmose, M. L. Rank, M. Ungstrup, K. Rosenkranz, and D. P. Mandic, "An in-the-ear platform for recording electroencephalogram," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2011, 6882–5.

[74] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, 4(2), 2007, R1, DOI: 10.1088/1741-2560/4/2/R01.

[75] J. H. McDermott, "The cocktail party problem," *Current Biology*, 19(22), 2009, R1024–R1027.

[76] A. Meiser and M. G. Bleichner, "Ear-EEG compares well to cap-EEG in recording auditory ERPs: a quantification of signal loss," *Journal of Neural Engineering*, 19(2), 2022, 026042.

[77] A. Meiser, F. Tadel, S. Debener, and M. G. Bleichner, "The sensitivity of ear-EEG: evaluating the source-sensor relationship using forward modeling," *Brain Topography*, 33(6), 2020, 665–76.

[78] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, 485(7397), 2012, 233.

[79] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex," *Journal of neurophysiology*, 102(6), 2009, 3329–39.

[80] B. Mirkovic, M. G. Bleichner, M. De Vos, and S. Debener, "Target speaker detection with concealed EEG around the ear," *Frontiers in Neuroscience*, 10, 2016, 349.

[81] 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*, 12(4), 2015, 046007.

[82] 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, 941–5.

[83] A. M. Narayanan and A. Bertrand, "Analysis of miniaturization effects and channel selection strategies for EEG sensor networks with application to auditory attention detection," *IEEE Transactions on Biomedical Engineering*, 67(1), 2019, 234–44.

[84] A. M. Narayanan, P. Patrinos, and A. Bertrand, "Optimal versus approximate channel selection methods for EEG decoding with application to topology-constrained neuro-sensor networks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29, 2020, 92–102.

[85] A. M. Narayanan, R. Zink, and A. Bertrand, "EEG miniaturization limits for stimulus decoding with EEG sensor networks," *Journal of Neural Engineering*, 18(5), 2021, 056042.

[86] W. Nogueira, G. Cosatti, I. Schierholz, M. Egger, B. Mirkovic, and A. Büchner, "Toward Decoding Selective Attention From Single-Trial EEG Data in Cochlear Implant Users," *IEEE Transactions on Biomedical Engineering*, 67(1), 2019, 38–49.

[87] W. Nogueira, H. Dolhopiatenko, I. Schierholz, A. Büchner, B. Mirkovic, M. G. Bleichner, and S. Debener, "Decoding selective attention in normal hearing listeners and bilateral cochlear implant users with concealed ear EEG," *Frontiers in Neuroscience*, 13, 2019, 720.

[88] 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*, 25(7), 2015, 1697–706.

[89] C. O'Callaghan, K. Kveraga, J. M. Shine, R. B. Adams Jr, and M. Bar, "Predictions penetrate perception: Converging insights from brain, behaviour and disorder," *Consciousness and cognition*, 47, 2017, 63–74.

[90] A. Parthasarathy, K. E. Hancock, K. Bennett, V. DeGruttola, and D. B. Polley, "Bottom-up and top-down neural signatures of disordered multi-talker speech perception in adults with normal hearing," *eLife*, 9, 2020, e51419, DOI: 10.7554/eLife.51419.

[91] J. Peelle and A. Wingfield, "How Our Brains Make Sense of Noisy Speech," *Acoustics Today*, 18(3), 2022, 40–8.

[92] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, 2018.

[93] M. K. Pichora-Fuller and G. Singh, "Effects of age on auditory and cognitive processing: implications for hearing aid fitting and audiologic rehabilitation," *Trends in amplification*, 10(1), 2006, 29–59.

[94] C. Puffay, J. Van Canneyt, J. Vanthornhout, T. Francart, *et al.*, "Relating the fundamental frequency of speech with EEG using a dilated convolutional network," *arXiv preprint arXiv:2207.01963*, 2022.

[95] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, 8(4), 2000, 441–6.

[96] S. A. Fuglsang, T. Dau, and J. Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, 156, 2017, 435–44.

[97] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12), 2017, 2257–71.

[98] J. Z. Simon, "Human auditory neuroscience and the cocktail party problem," in *The Auditory System at the Cocktail Party*, Springer, 2017, 169–97.

[99] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nature neuroscience*, 9(4), 2006, 578–85.

[100] L. Straetmans, B. Holtze, S. Debener, M. Jaeger, and B. Mirkovic, *"Neural Tracking to go"*, OpenNeuro, 2021, DOI: doi:10.18112/openneuro.ds003801.v1.0.0.

[101] E. Su, S. Cai, P. Li, L. Xie, and H. Li, "Auditory attention detection with EEG channel attention," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, 5804–7.

[102] 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*, 2022.

[103] S. R. Synigal, E. S. Teoh, and E. C. Lalor, "Including measures of high gamma power can improve the decoding of natural speech from EEG," *Frontiers in Human Neuroscience*, 14, 2020, 130.

[104] A. Taherkhani, A. Belatreche, Y. Li, G. Cosma, L. P. Maguire, and T. M. McGinnity, "A review of learning in biologically plausible spiking neural networks," *Neural Networks*, 122, 2020, 253–72.

[105] 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*, 51(5), 2020, 1234–41.

[106] P. Tremblay, V. Brisson, and I. Deschamps, "Brain aging and speech perception: Effects of background noise and talker variability," *NeuroImage*, 227, 2021, 117675.

[107] S. Tune, M. Alavash, L. Fiedler, and J. Obleser, "Neural attentional-filter mechanisms of listening success in middle-aged and older individuals," *Nature Communications*, 12(1), 2021, 1–14.

[108] J. J. Van Berkum, "The brain is a prediction machine that cares about good and bad-any implications for neuropragmatics?" *Italian Journal of Linguistics*, 22, 2010, 181–208.

[109] 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*, 64(5), 2016, 1045–56.

[110] 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*, 10, 2021, e56481.

[111] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in EEG signal analysis," *Neurocomputing*, 421, 2021, 1–14.

[112] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10), 2018, 1702–26.

[113] L. Wang, E. X. Wu, and F. Chen, "EEG-based auditory attention decoding using speech-level-based segmented computational models," *Journal of Neural Engineering*, 18(4), 2021, 046066.

[114] M. Wöstmann, B. Herrmann, B. Maess, and J. Obleser, "Spatiotemporal dynamics of auditory attention synchronize with speech," *Proceedings of the National Academy of Sciences*, 113(14), 2016, 3873–8.

[115] M. Wöstmann, B. Herrmann, A. Wilsch, and J. Obleser, "Neural alpha dynamics in younger and older listeners reflect acoustic challenges and predictive benefits," *Journal of Neuroscience*, 35(4), 2015, 1458–67.

[116] M. Wöstmann, J. Vosskuhl, J. Obleser, and C. S. Herrmann, "Opposite effects of lateralised transcranial alpha versus gamma stimulation on auditory spatial attention," *Brain Stimulation*, 11(4), 2018, 752–8.

[117] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016," *IEEE Transactions on Cognitive and Developmental Systems*, 14(1), 2020, 4–19.

[118] M. Yang, S. A. Sheth, C. A. Schevon, G. M. M. Ii, and N. Mesgarani, "Speech reconstruction from human auditory cortex with deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[119] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, 241–5.