COGNITIVE-DRIVEN SPEECH ENHANCEMENT USING EEG-BASED AUDITORY ATTENTION DECODING FOR HEARING AID APPLICATIONS

Von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
angenommene Dissertation

von

Ali Aroudi
geboren am 11. April 1987
in Esfahan (Iran)

Ali Aroudi: *Cognitive-driven speech enhancement using EEG-based auditory attention decoding for hearing aid applications*

ERSTGUTACHTER:
Prof. Dr. ir. Simon Doclo, *University of Oldenburg, Germany*

WEITERE GUTACHTER:
Prof. Dr. Bernd T. Meyer, *University of Oldenburg, Germany*
Prof. Dr. Alexander Bertrand, *KU Leuven, Belgium*

TAG DER DISPUTATION:
09. November 2020

ACKNOWLEDGMENTS

This thesis has been written at the Signal Processing Group at the Department of Medical Physics and Acoustics of the University of Oldenburg. I would like to take this opportunity to express my gratitude for the help and support of many people who have contributed to its completion.

I would like to thank my supervisor Simon Doclo for giving me the opportunity to write this thesis and providing me with support and valuable guidance to improve my work in many different ways through the years. I am grateful that he gave me the freedom to pursue my interests, while still providing numerous ideas. I would also like to thank Bernd Meyer and Alexander Bertrand for taking the time to read and evaluate my thesis.

This research was mainly funded by the cluster of excellence Hearing4all and the Signals and Cognition PhD program and being involved in these cluster and program was a great professional and personal experience. It gave me the opportunity to meet talented researchers and great people, and I benefited from the many interesting and fun discussions. Also, I would like to thank WS Audiology in Erlangen, NTT Communication Science Laboratories in Kyoto, and Microsoft in Munich and Redmond for giving me the internship opportunities, providing me with great guidance, and making my internship stays very memorable experiences. I am especially grateful to Maja Serman, Eghart Fischer, Henning Puder, Tomohiro Nakatani, Marc Delcroix, Shoko Araki, Ivan Tashev and Sebastian Braun.

It is my great pleasure to thank all the academic and administrative staff of the Department of Medical Physics and Acoustics, especially Katja Warnken and Jennifer Köllner. I am particularly grateful to all current and former members of the Signal Processing Group with whom we did this journey together while supporting and encouraging each other, especially Daniel Marquardt, Ante Jukić, Kai Siedenburg, Ina Kodrasi, Nico Gößling, Reza Varzandeh, Dörte Fischer, Marvin Tammen, Daniel Fejgin and Henning Schepker.

Finally, my greatest gratitude to my parents and to my brother for believing in me ever since I remember. Most importantly, I would like to thank my beloved Elham for her unlimited encouragement and the everlasting support.

Oldenburg, November 2020
Ali Aroudi

ABSTRACT

Identifying the target speaker in hearing aid applications is an essential ingredient to improve speech intelligibility. Although several speech enhancement algorithms are available to reduce background noise or to perform source separation in multi-speaker scenarios, their performance depends on correctly identifying the target speaker to be enhanced. Recent advances in electroencephalography (EEG) have shown that it is possible to identify the target speaker which the listener is attending to using single-trial EEG-based auditory attention decoding (AAD) methods. However, in realistic acoustic environments the AAD performance is influenced by undesired disturbances such as interfering speakers, noise and reverberation. In addition, it is important for real-world hearing aid applications to close the AAD loop by presenting on-line auditory feedback.

This thesis deals with the problem of identifying and enhancing the target speaker in realistic acoustic environments based on decoding the auditory attention of the listener using single-trial EEG recordings. To this end, we thoroughly analyze the AAD performance in noisy and reverberant environments, we propose novel methods for decoding auditory attention and we propose open-loop and closed-loop cognitive-driven speech enhancement systems for hearing aid applications.

First, we analyze the impact of different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy) on the performance of a least-squares-based AAD method. We show that for all considered acoustic conditions it is possible to decode auditory attention with a considerably large decoding performance, but that the decoding performance is significantly affected by the presence of background noise and especially the interfering speaker in the reference signals used for decoding.

Second, we propose several open-loop and closed-loop cognitive-driven speech enhancement systems. The first system is an open-loop cognitive-driven binaural beamformer, aiming at enhancing the target speaker and suppressing the interfering speaker and background noise while preserving the spatial impression of the acoustic scene. In this system a binaural minimum-variance-distortionless-response (MVDR) or binaural linearly-constrained-minimum-variance (LCMV) beamformer is steered based on AAD. For a two-speaker scenario in diffuse babble noise we show that the proposed cognitive-driven binaural beamforming system yields a significantly larger speech enhancement performance than a fixed forward-steered binaural MVDR beamformer, both in anechoic as well as reverberant conditions. The second system is an open-loop cognitive-driven convolutional beamformer, aiming at enhancing the target speaker and jointly suppressing the interfering speaker, reverberation and background noise. This system combines a neural-network-based mask estimator, convolutional beamformers and AAD. We show that the proposed cognitive-driven convolutional beamforming system yields a significantly larger speech enhancement

performance than cognitive-driven systems based on conventional beamformers. The third system is a closed-loop cognitive-driven gain controller, where real-time AAD enables the listener to directly interact with an adaptive gain controller. Although there is a significant delay to detect attention switches, experimental results demonstrate the feasibility of the proposed system, which is able to improve the SIR between the attended and the unattended speaker.

Third, we propose novel methods to decode auditory attention. More specifically, we propose a reference signal generation approach based on binary masking, which uses binary masks based on directional speech presence probability to discard low-energy intervals which are susceptible to interfering speech and background noise. In addition, we propose an AAD method based on a state-space model, which translates the correlation coefficients into more reliable probabilistic attention measures and improves the decoding performance of linear and non-linear methods using small correlation windows.

ZUSAMMENFASSUNG

Die Bestimmung des Zielsprechers in Hörgeräte-Anwendungen ist ein wesentlicher Bestandteil zur Verbesserung der Sprachverständlichkeit. Obwohl verschiedene Sprachverbesserungsalgorithmen zur Reduzierung von Hintergrundgeräuschen oder zur Quellentrennung in Szenarien mit mehreren Sprechern existieren, hängt ihre Leistung von der korrekten Bestimmung des zu verbessernden Zielsprechers ab. Jüngste Entwicklungen in der Elektroenzephalographie (EEG) haben gezeigt, dass es möglich ist, den Zielsprecher, auf den sich der Hörer konzentriert, mit Hilfe von AAD-Methoden (Auditive Aufmerksamkeitsdekodierung), die auf einem einzelnen Durchlauf im EEG basieren, zu bestimmen. In realistischen akustischen Umgebungen wird die Leistung der AAD jedoch durch unerwünschte Störfaktoren wie interferierende Sprecher, Geräusch und Nachhall beeinträchtigt. Darüber hinaus ist es für die Anwendung von Hörgeräten in der realen Welt wichtig, den AAD-Kreislauf zu schließen, indem online ein auditives Feedback dargeboten wird.

Diese Arbeit befasst sich mit dem Problem der Bestimmung und Verstärkung des Zielsprechers in realistischen akustischen Umgebungen, basierend auf der Dekodierung der auditiven Aufmerksamkeit des Hörers unter Verwendung einzelner EEG-Aufnahmen. Zu diesem Zweck analysieren wir eingehend die Leistung der AAD in verrauschten und nachhallenden Umgebungen, wir präsentieren neue Methoden zur Dekodierung der auditiven Aufmerksamkeit und wir präsentieren kognitiv gesteuerte Sprachverbesserungssysteme mit offenem und geschlossenem Regelkreis für Hörgeräteanwendungen.

Als erstes analysieren wir den Einfluss verschiedener akustischer Bedingungen (anechoisch, nachhallend, verrauscht und nachhallend-verrauscht) auf die Leistung einer auf den kleinsten Quadraten basierenden AAD-Methode. Wir zeigen, dass es für alle betrachteten akustischen Bedingungen möglich ist, die auditive Aufmerksamkeit mit einer beträchtlich großen Dekodierleistung zu dekodieren, dass aber die Dekodierleistung durch das Vorhandensein von Hintergrundgeräuschen und insbesondere den interferierenden Sprecher in den zur Dekodierung verwendeten Referenzsignalen signifikant beeinflusst wird.

Als zweites präsentieren wir mehrere kognitiv gesteuerte Sprachverbesserungssysteme mit offenem und geschlossenem Regelkreis. Das erste System ist ein kognitiv gesteuerter binauraler Beamformer mit offenem Regelkreis, der darauf abzielt, den Zielsprecher zu verstärken und den interferierenden Sprecher und Hintergrundgeräusche zu unterdrücken, während der räumliche Eindruck der akustischen Szene erhalten werden soll. In diesem System wird ein binauraler MVDR-Beamformer (Minimum-Variance-Distortionless-Response) oder ein binauraler LCMV-Beamformer (Linearly-Constrained-Minimum-Variance) auf der Basis von AAD gesteuert. Für ein Zwei-Sprecher-Szenario mit diffusem Sprachgeräusch

zeigen wir, dass das vorgestellte kognitiv-gesteuerte binaurale Beamformer-System sowohl unter anechoischer als auch unter nachhallender Bedingung eine signifikant größere Sprachverbesserungsleistung erzielt als ein fixer nach vorne gesteuerter binauraler MVDR-Beamformer. Das zweite System ist ein kognitiv gesteuerter Faltungs-Beamformer mit offenem Regelkreis, der darauf abzielt, den Zielsprecher zu verstärken und gleichzeitig den interferierenden Sprecher, den Nachhall und das Hintergrundgeräusch zu unterdrücken. Dieses System kombiniert einen auf neuronalen Netzwerken basierenden Maskenschätzer, Faltungs-Beamformer und AAD. Wir zeigen, dass das vorgeschlagene kognitiv-gesteuerte Faltungs-Beamformersystem eine signifikant größere Sprachverbesserung erzielt, als kognitiv gesteuerte Systeme, die auf konventionellen Beamformern basieren. Das dritte System ist ein kognitiv gesteuerter Verstärkungsregler mit geschlossenem Regelkreis, bei dem die Echtzeit-AAD dem Hörer die direkte Interaktion mit einem adaptiven Verstärkungsregler ermöglicht. Obwohl es eine erhebliche Zeitverzögerung bei der Erkennung von Aufmerksamkeitswechseln gibt, zeigen die experimentellen Ergebnisse die Umsetzbarkeit des vorgeschlagenen Systems, das in der Lage ist, den SIR zwischen dem Sprecher, auf den sich der Hörer konzentriert und dem Sprecher, auf den sich der Hörer nicht konzentriert, zu verbessern.

Als drittes schlagen wir neue Methoden zur Dekodierung der auditiven Aufmerksamkeit vor. Konkret schlagen wir einen auf binärer Maskierung basierenden Ansatz zur Generierung von Referenzsignalen vor, bei dem binären Masken auf der Grundlage direktionaler Sprachanwesenheitswahrscheinlichkeit verwendet werden, um Intervalle mit niedriger Energie zu verwerfen, die besonders anfällig für interferierende Sprache und Hintergrundgeräusche sind. Darüber hinaus schlagen wir eine auf einem Zustandsraummodell basierende AAD-Methode vor, die die Korrelationskoeffizienten in zuverlässigere probabilistische Aufmerksamkeitsmaße übersetzt und die Dekodierungsleistung linearer und nichtlinearer Methoden unter Verwendung kleiner Korrelationsfenster verbessert.

GLOSSARY

Acronyms and abbreviations

AAD	auditory attention decoding
AGC	adaptive gain controller
BCI	brain-computer interface
BBEAM	binaural beamformer
BEAM	beamformer
BLSTM	bi-directional long short term memory network
cGMM	complex Gaussian mixture model
CASA	computational auditory scene analysis
CNN	convolutional neural network
DNN	deep neural network
DOA	direction of arrival
DSPP	directional speech presence probability
ECoG	electrocorticographic
EEG	electroencephalography
EM	expectation-maximization
ERP	event-related potential
fwSSNR	frequency-weighted segmental SNR
GCC-PHAT	generalized cross-correlation with phase transform
GLM	generalized linear model
IC	interaural coherence
ILD	interaural level difference
ISTFT	inverse short-time Fourier transform
ITD	interaural time difference
LSL	Lab Streaming Layer
LTI	linear time-invariant
LCMP	linearly constrained minimum power
LCMV	linearly-constrained-minimum-variance
MAP	maximum posterior distribution

MMSE	minimum mean-squared error
MPDR	minimum power distortionless response
MVDR	minimum-variance-distortionless-response
MWF	multi-channel Wiener filter
PSD	power spectral density
RAAD	real-time AAD
RETF	relative early transfer function
RTF	relative transfer function
RMS	root-mean-square
SINR	signal-to-interference-plus-noise ratio
SIR	signal-to-interference ratio
SNR	signal-to-noise ratio
SPP	speech presence probability
SRM	spatial release from masking
SSM	state-space model
STFT	short-time Fourier transform
SVM	support vector machine
TRF	temporal response function
wLCMP	weighted linearly-constrained-minimum-power
wMPDR	weighted minimum-power-distortionless-response
WOLA	weighted overlap-add

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

1.1 Motivation

Hearing impairment is affecting about half a billion persons in the world, including every third person aged 65 [2]. Hearing impairment often leads to a decreased speech intelligibility in complex acoustic conditions, particularly in multi-speaker scenarios. Persons with impaired hearing typically have serious difficulties to segregate a target speaker from a mixture of speakers and background noise, reducing their ability to interactively communicate with other persons and possibly leading to feelings of isolation [3,4] and cognitive decline [5,6]. Hearing aids aim at restoring the normal hearing ability by several processing steps such as frequency-dependent amplification, dynamic range compression and speech enhancement [7,8]. The main objective of speech enhancement is to improve the intelligibility of the recorded microphone signals, which are often corrupted by various undesired disturbances, such as interfering speakers, noise and reverberation. During the last decades many contributions have been dedicated to speech enhancement for hearing aid applications, and several single- and multi-channel noise reduction methods have been proposed [9–15]. Typically, the performance of these methods in terms of improving speech intelligibility depends on correctly identifying the target speaker to be enhanced. In hearing aid applications, the target speaker is often assumed to be either located in front of the hearing aid user or to be the loudest speaker. However, since in real-world conditions these assumptions are often violated, the performance of speech enhancement methods may substantially decrease. Hence, correctly identifying the target speaker in hearing aid applications is an essential ingredient to successfully improve speech intelligibility, motivating the work in this thesis.

Recent advances have shown that it is possible to infer the auditory attention of listeners from electroencephalography (EEG) recordings [16, 17]. Using single-trial EEG recordings, several auditory attention decoding (AAD) methods have been proposed to identify the attended speaker [16, 18–26]. The possibility of decoding auditory attention from EEG recordings has led to an increasing research interest to incorporate AAD in a brain-computer interface (BCI) for real-world applications, e.g., to cognitively control speech enhancement in hearing aids [27–31]. Cognitive-driven speech enhancement systems using AAD potentially provide the listener with a relatively high degree of flexibility to selectively attend to a desired speaker.

However, in realistic acoustic environments the performance of AAD is likely to be influenced by undesired disturbances (interfering speakers, noise and reverberation), which may have a negative effect on the reference signals used in AAD methods [19, 26, 27, 32]. Additionally, it is important to close the loop between the cognitive-driven speech enhancement system and the listener by presenting the enhanced speech signal in an on-line fashion, enabling the listener to interact with the speech enhancement system. **The main objectives of this thesis are therefore to analyze the performance of AAD in realistic noisy and reverberant acoustic conditions, to improve AAD methods and to develop and evaluate open-loop and closed-loop cognitive-driven speech enhancement systems for hearing aid applications using AAD.**

In the remainder of this chapter we present a brief overview of state-of-the-art speech enhancement and AAD methods and highlight the main contributions of this thesis. In Section 1.2, we discuss the effect of reverberation, noise and interfering speakers on speech intelligibility and cues. In Section 1.3, we provide a general overview of single- and multi-channel speech enhancement methods and discuss some remaining challenges of applying these methods in hearing aids. In Section 1.4, we provide an overview of AAD methods and discuss the main challenges of applying theses methods for hearing aid applications. In Section 1.5, we provide an overview of open-loop cognitive-driven speech enhancement systems using AAD and discuss some remaining challenges of these systems for hearing aid applications. In Section 1.6, we summarize the structure and the main contributions of the thesis.

1.2 Effect of reverberation, noise and interfering speaker on speech intelligibility and cues

Throughout the thesis, we typically consider an acoustic scenario comprising two competing speakers and background noise in a reverberant environment, as illustrated in Fig. 1.1. The listener is wearing binaural hearing aids with multiple microphones. The signals at ears of the listener and the signals captured by the hearing aid microphones consist of a mixture of the target speaker, the interfering speaker, background noise and reverberation. The target speaker is defined as the speech source to which the listener wants to listen, whereas the interfering speaker is the speech source to which the listener does not want to listen and should hence be suppressed. Typical background noise can be, e.g., fan noise, traffic noise or babble noise which occurs when multiple speakers are talking simultaneously. Background noise has a negative effect on speech intelligibility, even more for hearing-impaired listeners [33–36].

In a reverberant environment, the speech signals of the target speaker and the interfering speaker are not only directly arriving at the ears of the listener and the hearing aid microphones, but are also acoustically reflected against surfaces and objects. In order to consider the influence of reverberation on speech intelligibility, reverberation is typically decomposed into early and late reverberation. Early reverberation consists of spatially distinct reflections [37], which arrive at the ears and the hearing aid microphones typically in the order of tens of milliseconds [37],

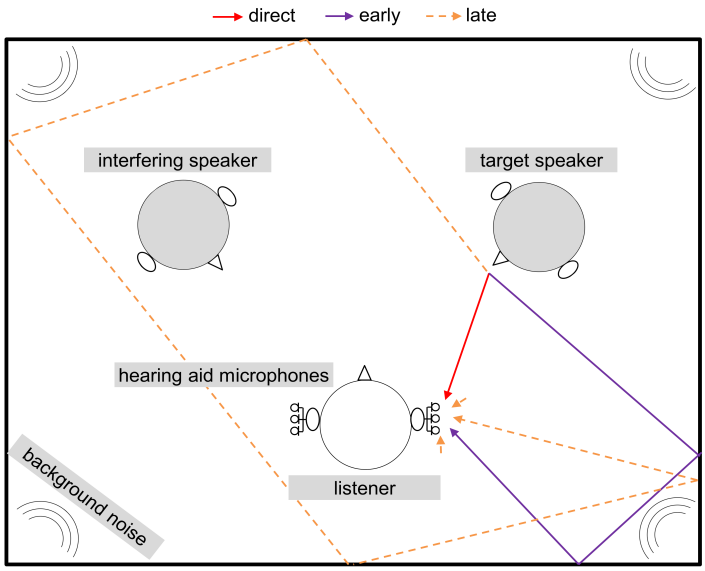


Fig. 1.1: Acoustic scenario with a target speaker, an interfering speaker and background noise in a reverberant environment. The listener is wearing binaural hearing aids with multiple microphones. The direct path, early and late reverberation are illustrated between the target speaker and the microphones of the right hearing aid.

whereas late reverberation occurs later and arrives approximately uniformly from all directions [37]. While early reflections can be beneficial for speech intelligibility by increasing the efficient signal-to-noise ratio of the target speech [38–40], late reverberation is known to have a detrimental effect on speech quality and intelligibility [39, 41, 42], in particular for hearing-impaired listeners [42]. The amount of reverberation in a room can be specified using the reverberation time (T_{60}) [43], i.e., the time required for the reverberant energy to decay by 60 dB after the sound source has been deactivated.

In complex acoustic conditions, the human auditory system possesses a remarkable ability to segregate a speaker of interest from a mixture of speakers and background noise. While the auditory system using one ear, referred to as monaural hearing, may perform source segregation based on monaural cues (e.g., pitch, spectral pinna cues and amplitude modulation), using both ears, referred to as binaural hearing, allows the auditory system to exploit binaural cues in addition to monaural cues for segregation.

The main binaural cues are the interaural level difference (ILD), the interaural time difference (ITD) and the interaural coherence (IC). The ILD occurs due to the head acting as an obstacle for the sound waves traveling between the ears, i.e., the so-called head shadow effect. The ITD occurs due to the time difference of arrival of the sound waves traveling between the ears. The interaural coherence is the

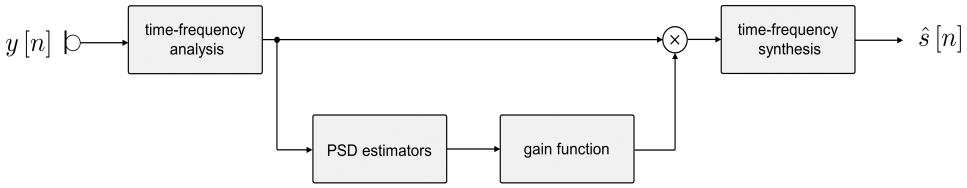


Fig. 1.2: A block scheme of a typical single-channel speech enhancement system. $y[n]$ denotes the captured microphone signal and $\hat{s}[n]$ denotes the estimated target speech signal.

normalized cross-correlation between the signals at both ears. It has been shown that the ILD cue dominantly contributes to source localization at high frequencies [44,45], whereas the ITD cue dominantly contributes at low frequencies. These binaural cues together play an important role in speech intelligibility [45–47] by providing spatial release from masking (SRM), also known as binaural unmasking [45–47]. The SRM is relatively small when the target speech and the interfering speech can be segregated by monaural cues, e.g., when the speakers have a different pitch, and is particularly large when other cues, e.g., binaural cues, are available for the listener. In noisy and reverberant environments the binaural cues are smeared such that source localization becomes more difficult and SRM is reduced, particularly for impaired-hearing persons [46, 48, 49].

1.3 Overview of speech enhancement methods

The aim of speech enhancement is to selectively enhance the target speech signal and suppress interfering speakers and background noise. In general, speech enhancement methods can be classified into single-channel and multi-channel methods, based on the number of microphones which are used. In the following sections, we present an overview of several single- and multi-channel speech enhancement methods. A more detailed review of speech enhancement methods can be found in, e.g., [10–15]. In addition, we present a general overview of multi-channel speech enhancement methods for binaural hearing aid applications and discuss challenges and open issues.

1.3.1 *Single-channel speech enhancement*

Single-channel speech enhancement methods typically exploit spectro-temporal differences between the speech and the noise signals. The spectral coefficients of the target speech signal are commonly estimated by applying a (real-valued) gain to the spectral coefficients of the noisy microphone signal, aiming at suppressing the disturbances present in the recorded microphone signal. A block diagram of a typical single-channel speech enhancement system is depicted in Fig. 1.2. Many gain functions have been derived using a statistical framework, e.g., resulting in estimators

that are optimal in the minimum mean-squared error (MMSE) or the maximum posterior distribution (MAP) sense under certain assumptions. The well-known Wiener filter is the MMSE estimator of the target speech spectral coefficients, when assuming a Gaussian distribution for the speech and noise spectral coefficients [50–53]. However, since a super-Gaussian assumption is more appropriate to model speech spectral coefficients, several MMSE and MAP estimators using a super-Gaussian assumption have been proposed [15, 54–58].

To compute these gain functions, typically an estimate of the short-term power spectral densities (PSDs) of the speech signal and the noise is required, as shown in Fig. 1.2. Several methods have been proposed to blindly estimate the noise PSD from the noisy microphone signals, e.g., based on a voice activity detector [59], speech presence probability (SPP) [60,61], or by using subspace techniques [62]. In addition, several methods have been proposed to estimate the speech PSD, e.g., the maximum likelihood estimation method in [63] and the temporal cepstrum smoothing method in [64].The aforementioned PSD estimators typically suffer from a limited capability to track highly non-stationary background noises, e.g., resulting in speech distortion and limited noise reduction [52, 53]. Alternatively, the noise and speech PSDs can be estimated using machine learning methods, where the structure and statistics of speech and noise are learned and modeled using training data. Several machine learning-based PSD estimation methods have been proposed by using statistical generative models, e.g., hidden Markov models [58, 65–68], by using probabilistic models employing non-negative matrix factorization [69–71], and by using codebook-based methods [72–74].

More recently, several single-channel speech enhancement methods based on neural networks have been proposed. Neural networks have the potential to approximate arbitrary functions [75,76], referred to as the universal approximation property. Neural networks with many hidden layers, known as deep neural networks (DNNs), can be trained to directly estimate the target speaker [77–79]. In addition, DNNs can be trained to estimate gain functions [80–82] or masks, aiming at clustering and retaining the time-frequency bins containing mainly the target speaker [83–85]. DNNs with different structures have been considered for speech enhancement, e.g., recurrent neural networks [77, 78, 85] and convolutional neural networks (CNNs) [79, 83, 84]. It should be noted that the generalization of the neural network-based speech enhancement methods to unseen conditions, e.g., unknown speakers or unseen acoustic conditions, is still a challenging problem.

1.3.2 *Multi-channel speech enhancement*

When multiple microphones are available multi-channel speech enhancement allows to exploit the spatial diversity between the target speaker, the interfering speakers and the background noise in addition to the spectro-temporal diversity. Spatial filters, also often referred to as beamformers [9], can preserve the target speech signal without distortion (or with low distortion) and suppress the interfering speech signals and background noise by linearly filtering and summing the microphone signals (see Fig. 1.3). Generally, beamformers can be categorized into two main classes, i.e.,

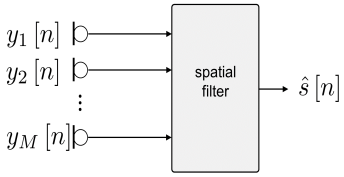


Fig. 1.3: A block scheme of spatial filtering. $y_m[n]$ denotes the m-th captured microphone signal with $m \in \{1, \cdots, M\}$ and $\hat{s}[n]$ denotes the estimated target speech signal.

signal-independent and signal-dependent, based on whether the statistics of the microphone signals are exploited or not [13]. In the following section, we present an overview of the existing signal-independent and signal-dependent beamformers proposed in the literature. In addition, we present an overview of the existing methods for estimating the parameters of these beamformers.

Signal-independent beamformer

Signal-independent beamformers, also referred to as fixed beamformers, require the geometry of the microphone array and the direction of arrival (DOA) of the target speaker to be known. The most commonly used signal-independent beamformers are the delay-and-sum beamformer, the superdirective beamformer and the differential microphone beamformer. The delay-and-sum beamformer simply aligns the signals arriving from the DOA of the target speaker by applying delays to the microphone signals before summing them, thereby maximizing the array gain for spatially uncorrelated noise, e.g., sensor noise [9]. The superdirective beamformer maximizes the array gain assuming that the noise field is diffuse (or spatially isotropic), i.e., composed of a superposition of uncorrelated plane waves that are uniformly distributed on a surface with equal PSDs [86]. The differential microphone beamformer is implemented using small-sized microphone arrays [87, 88].

Most aforementioned signal-independent beamformers require the DOAs of speakers. The DOA of the target speaker can be estimated, e.g., by using a beamforming-based method [89] or by using a classification-based SPP estimation method [90].

In general, since the statistics of the microphone signals are not taken into account in the design of signal-independent beamformers, these beamformers are not able to adapt to time-varying acoustic environments, e.g., when opening a window or turning on the air conditioner.

Signal-dependent beamformer

In contrast to signal-independent beamformers, signal-dependent beamformers exploit the spectro-temporal and spatial statistics of the microphone signals. The most commonly used signal-dependent beamformers exploiting the spatial statistics are the minimum variance distortionless response (MVDR) beamformer and the linearly constrained minimum variance (LCMV) beamformer. The commonly used signal-

dependent beamformer exploiting both the spectro-temporal and spatial statistics is the multi-channel Wiener filtering (MWF).

The MVDR beamformer aims at minimizing the PSD of the output signal while preserving the target speech component in an arbitrarily chosen reference microphone [9, 10, 86]. The MVDR beamformer requires either the covariance matrix of the microphone signals, the covariance matrix of the interfering speech and background noise component, referred to as the covariance matrix of the overall interfering component, or the covariance matrix of the background noise component. In addition, the MVDR beamformer requires the estimate of the reverberant relative transfer function (reverberant RTF) between the target speaker and the microphones. From a theoretical point of view, if a perfect estimate of the covariance matrix of either the microphone signals or the overall interfering component and a perfect estimate of the reverberant RTF of the target speaker are used in designing the MVDR beamformer, the MVDR beamformer is optimal in terms of signal-to-interference-plus-noise (SINR) ratio. The MVDR beamformer using an estimate of the covariance matrix of the microphone signals is also known as minimum power distortionless response (MPDR) beamformer [9]. In practice, it should be realized that using the estimated covariance matrix of the microphone signals may lead to the target speaker cancellation in the case of reverberant RTF estimation errors [9, 91]. If perfect estimates of the covariance matrix of the background noise component and the reverberant RTF of the target speaker are used, the MVDR beamformer is optimal in terms of signal-to-noise ratio (SNR).

In order to also control the interfering speaker suppression, in [9] an extension of the MVDR incorporating an interference suppression constraint has been proposed, known as LCMV beamformer. The LCMV beamformer requires either the covariance matrix of the microphone signals, the covariance matrix of the overall interfering component or the covariance matrix of the background noise component. In addition, the LCMV beamformer requires the estimates of the reverberant RTFs of the target speaker and the interfering speakers. From a theoretical point of view, if perfect estimates of the RTFs and one of the mentioned covariance matrices are used and the interference suppression constraint is set to completely suppress the interfering speakers, the LCMV beamformer is optimal in terms of signal-to-interference (SIR) ratio. The LCMV beamformer using an estimate of the covariance matrix of the microphone signals is known as linearly constrained minimum power (LCMP) beamformer [9].

The MWF generates an MMSE estimate of the target speech component at an arbitrarily chosen reference microphone [92, 93]. The MWF requires the covariance matrix of the target speech component and of the overall interfering component. It has been shown that for a single speech source the MWF can be decomposed as an MVDR beamformer followed by a single-channel Wiener filter [94, 95]. As opposed to the MDVR and the LCMV beamformer, at the output of the MWF there is a trade-off between distortion of the target speech component and reduction of the interfering speech and noise component. To control this trade-off, the MWF has been extended to a speech-distortion-weighted MWF [96, 97].

While the discussed signal-dependent beamformers are able to suppress interfering speakers and background noise, they are not designed to suppress late reverberation, which has a negative effect on speech quality and intelligibility [39, 41, 42]. Aiming at preserving the target speech component in an arbitrarily chosen reference microphone and jointly suppressing the interfering speech component, the background noise component and the late reverberation component, a convolutional beamformer has been recently proposed in [98–100]. The convolutional beamformer can be implemented as a dereverberation filter followed by a variant of MPDR beamformer, referred to as weighted MPDR (wMPDR) beamformer [100]. This sequential implementation allows to estimate the parameters of the wMPDR beamformer using multi-channel dereverberated signals which are obtained as the outputs of the dereverberation filter. The convolutional beamformer requires the covariance matrix of microphone signals and the relative early transfer function (RETF) between the target speaker and the microphones.

All discussed beamformers require estimates of certain parameters, e.g., the covariance matrix of the microphone signals, the covariance matrix of the overall interfering component, the covariance matrix of the background noise component, or the covariance matrix of the target speaker component. In addition, the reverberant RTFs and the RETFs of the target speaker and the interfering speakers may be required. The covariance matrix of the microphone signals can be easily obtained by recursively updating microphone covariance matrix at each time-frequency (T-F) bin [13]. The covariance matrix of the overall interfering component is typically estimated by recursively updating covariance matrix of the microphone signals at the T-F bins where the target speaker is inactive [13], using the masks of the target speaker. The masks of the target speaker and also the interfering speakers can be estimated based on, e.g., the speech presence probabilities (SPPs) of the speakers at each T-F bin, estimated from the microphone signals. In [90] a classification-based directional SPP estimation method has been proposed using support vector machines (SVMs). In [101] a SPP estimation method has been proposed by modeling directional and sparse SPPs of the speakers for each T-F bin using a complex Gaussian mixture model (cGMM). The masks of speakers can be also directly estimated from the microphone signals using DNNs with different structures, e.g., bi-directional long short term memory networks (BLSTMs) and CNNs [83, 84, 102, 103].

The covariance matrix of the target speaker component can be estimated by subtracting the estimated covariance matrix of the microphone signals and the estimated covariance matrix of the overall interfering component [104]. Alternatively, the covariance matrix of the target speaker component can be estimated based on the generalized eigenvalue decomposition of the estimated covariance matrix of the microphone signals and the estimated covariance matrix of the overall interfering component [104].

The covariance matrix of the background noise component can be estimated by recursively updating covariance matrix of the microphone signals at the T-F bins where all speakers are inactive, using the masks of all speakers. Alternatively, the covariance matrix of the background noise component can be easily approximated using diffuse noise field assumption [11].

The reverberant RTFs of the target speaker and the interfering speakers are commonly estimated either using the covariance subtraction method [104–106] or the covariance whitening method [104, 107, 108]. Both of these RTF estimation methods are based on the rank-one model of the covariance matrix of the speaker. Let us consider as an example how the reverberant RTF of the target speaker can be estimated. The covariance subtraction method estimates the reverberant RTF of the target speaker based on the the estimated covariance matrix of the target speech component, which is obtained by subtraction of the covariance matrix of the microphone signals and the covariance matrix of the overall interfering component. Although the subtraction method has a relatively low computational complexity, its performance is not always very good since due to estimation errors the estimated covariance matrix of the target speech component typically does not have rank-one [104, 106]. The covariance whitening method estimates the reverberant RTF of the target speaker by applying the generalized eigenvalue decomposition either to the estimated covariance matrix of the microphone signals or the estimated covariance matrix of the target speech and background noise component, which can be obtained using the estimated masks of the target speaker. It should be noted that for an acoustic scenario with multiple simultaneously active speakers, due to the estimation errors of the covariance matrices the performance of jointly estimating the reverberant RTFs of the speakers can drastically decrease. As an alternative to reverberant RTF estimation, the RTFs of the speakers can be approximated by anechoic RTFs which are determined by the DOAs of the speakers. These anechoic RTFs can be either analytically computed based on a head model, e.g., [109], or selected from a database of (measured) prototype RTFs, e.g., [110]. However, using anechoic RTFs for speech enhancement in a reverberant environment may lead to a limited suppression of the interfering speakers.

Similarly to the reverberant RTF estimation, the RETFs of speakers are estimated using either the covariance subtraction method or the covariance whitening method [99, 100]. However, the covariance matrices used for RETF estimation are obtained using the multi-channel output signals of a dereverberation filter.

In general, since multi-channel speech enhancement exploits spatial diversity, it is typically able to yield a larger speech quality and speech intelligibility and a lower speech distortion compared to single-channel speech enhancement.

1.3.3 *Multi-channel speech enhancement in binaural hearing aids*

In order to enable hearing aid users to exploit binaural cues, the multi-channel speech enhancement methods discussed in Section 1.3.2 need to be extended to produce two different output signals, presented to the left and the right ear. In addition to reducing the background noise and the interfering speakers, another important objective of binaural speech enhancement methods is to preserve the binaural cues of the target speaker, the interfering speakers and the background noise. To generate two different outputs, all microphone signals from the left and right hearing aids are processed by two different (complex-valued) beamformers. In [95, 111, 112] the MVDR beamformer and the MWF have been extended into a binaural version

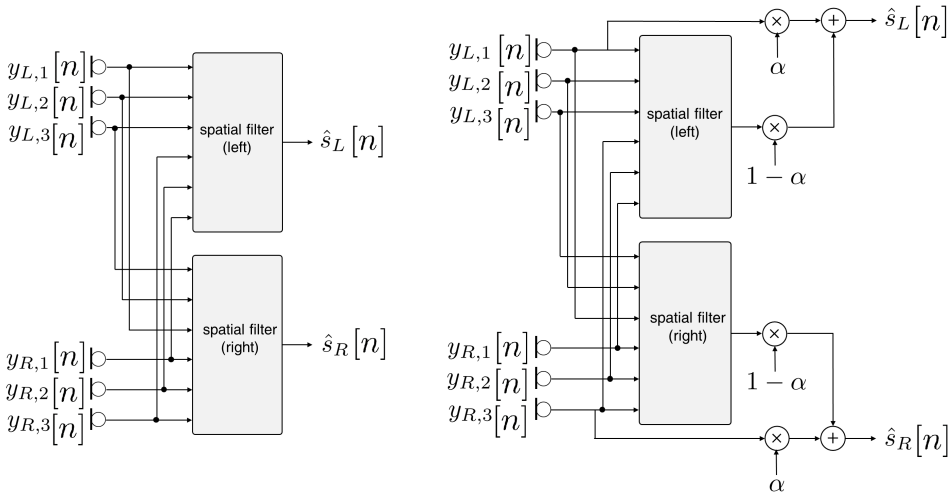


Fig. 1.4: Block scheme of binaural spatial filtering (a) incorporating constraints or (b) mixing with scaled noisy reference microphone signals. $y_{L,m}[n]$ and $y_{R,m}[n]$ denote the m-th captured microphone signal of the left and the right hearing aid with $m \in \{1,2,3\}$, $\hat{s}_L[n]$ and $\hat{s}_R[n]$ denote the estimated target speech signals of the left and the right hearing aid and α denotes the mixing parameter.

by estimating the target speech component at a reference microphone on the left and the right hearing aid. While the binaural MVDR beamformer and the binaural MWF preserve the binaural cues of the target speaker, these beamformers distort the binaural cues of the interfering speakers and the background noise, such that all sources are perceived as coming from the direction of the target speaker [95,113]. This binaural cue distortion may change the spatial impression of the acoustic scene, lead to a confusion between acoustical and visual information and decrease speech intelligibility [114].

In order to preserve the binaural cues of all sources in the acoustic scene, it has been proposed to apply a common gain to the reference microphone signals from both hearing aids, known as binaural spectral post-filtering technique [115]. Similarly to the single-channel speech enhancement methods discussed in Section 1.3.1, binaural spectral post-filtering generally aims at retaining the T-F bins containing mainly the target speaker by applying a large common gain and suppressing the T-F bins containing mainly interfering speakers or background noise by applying a small common gain. The common gain can be computed, e.g., by exploiting the spatial information between the microphones [116–118], based on computational auditory scene analysis (CASA) [119,120] or based on the output of a multi-channel speech enhancement algorithm [121]. Although binaural spectral post-filtering allows for binaural cue preservation of the target speaker, the interfering speakers and the background noise, it may introduce speech distortion, especially at low SNRs [51].

As an alternative to binaural spectral post-filtering, it has been proposed to either directly incorporate additional constraints into the binaural beamformer de-

sign [112,122,123] or to mix the binaural output signals with scaled noisy reference microphone signals [95,124–127], as depicted in Fig. 1.4. In [122] an extension of the binaural MVDR incorporating an interference suppression constraint has been proposed, known as the binaural LCMV beamformer. The binaural LCMV beamformer aims at partially suppressing the interfering speakers while perfectly preserving the binaural cues of both the target speaker and the interfering speakers. In addition, an extension of the binaural MVDR and the binaural MWF has been proposed by incorporating an RTF preservation constraint for the interfering speakers [112,123]. In [124] a binaural MVDR beamformer with partial noise estimation has been proposed, which mixes the output signals of the binaural MVDR beamformer with scaled noisy reference microphone signals using a mixing parameter determined based on psychoacoustically motivated boundaries. Furthermore, a binaural MWF with partial noise estimation has been proposed using a mixing parameter which is either frequency-independent [125] or determined based on psychoacoustically motivated boundaries [126] and or the output SNR [127]. It should be noted that mixing the output signals of either the binaural MVDR beamformer or the binaural MWF with scaled noisy reference microphone signals obviously results in a trade-off between preserving the binaural cues and reducing interfering speakers and background noise [95,124].

Challenges and open issues

The performance of most discussed binaural multi-channel speech enhancement methods depends on correctly estimating certain parameters, e.g., covariance matrices and reverberant RTFs. Although several RTF estimation methods are available for a single-speaker scenario [104,105,107], jointly estimating the RTFs of two simultaneously active speakers is not straightforward. Therefore, one goal of this thesis is to jointly estimate the RTFs or the RETFs of the speakers, either directly from the microphone signals based on the estimated DOAs of the speakers or from the covariance matrices based on the estimated masks of the speakers.

The wMPDR convolutional beamformer optimally combines interfering speaker suppression, dereverberation and noise suppression. While suppressing the interfering speaker is desired to improve speech intelligibility, keeping the interfering speaker audible is also important to allow the listener to switch attention between speakers. Therefore, another goal of this thesis is to develop an extension of the wMPDR convolutional beamformer, allowing to control the level of suppression of the interfering speaker.

Even when assuming that the parameters (covariance matrices and RTFs) can be accurately estimated and the binaural cues of all sources can be preserved, the performance in terms of speech intelligibility improvement still relies on correctly identifying the target speaker and the interfering speaker. In practice, this may not be successfully achieved by assuming that the target speaker is, e.g., located in front of the listener or is the loudest speaker. Therefore, another goal of this thesis is to identify the target speaker and the interfering speaker by decoding the auditory attention of the listener and subsequently enhancing the identified target speaker and

suppressing the identified interfering speaker using binaural multi-channel speech enhancement algorithms.

1.4 Overview of auditory attention decoding methods

The aim of auditory attention decoding (AAD) is to identify to which speaker the listener is attending by relating the EEG responses of the listener to the received speech signals. Using single-trial EEG recordings, several methods for AAD have been proposed to identify the attended speaker in an acoustic scenario with two competing speakers [16, 18, 19, 21–24, 128–132]. Existing AAD methods can be generally classified into forward-model-based and backward-model-based methods, depending on whether EEG recordings are predicted from speech envelopes or speech envelopes are predicted from EEG recordings. In the following sections, we briefly describe EEG signals, present an overview of forward-model-based and backward-model-based AAD methods and discuss challenges and open issues. A more detailed review of several AAD methods can be found in, e.g., [22, 128, 133].

1.4.1 *EEG*

The EEG represents the electrical potential, usually the difference in potential measured at various points on the scalp [134–137]. The electrical potential on the scalp is a superposition of neural activity-related current sources that are distributed in a volume conductor (the head). At the cellular level, neural synaptic activities (both excitatory as well as inhibitory) lead to local changes in the membrane potential [137]. A synchronized activity of many neuronal sets with a similar polarity constitutes a substantial current source. Different brain regions and states generate different sets of current sources. These current sources are mixed and constitute electrical potentials in the brain, referred to as neuronal action potentials or spikes, on the cortical surface, referred to as electrocorticographic (ECoG) activity, or on the scalp, referred to as EEG activity, as illustrated in Fig. 1.5.

Spike and ECoG measurements typically have a high spatial resolution and cover a wide frequency range. However, theses measurements are invasive. On the other hand, EEG measurements are non-invasive and have appropriate temporal resolution, e.g., to measure event-related potentials evoked by visual or auditory stimuli. Due to the non-invasive and appropriate temporal resolution properties of EEG measurements, EEG measurements have become an established measurement method for real-world applications, e.g., for clinical care and treatment as well as for BCI applications [134, 138–141]. Nevertheless, it should be realized that EEG measurements can be corrupted by environmental noise sources, such as AC power lines and mobile phones, and physiological noise sources, such as muscle artifacts caused by muscle contraction, eye movement and skin potentials. These noise sources may be effectively mitigated either by preventing noise during EEG recording or by removing noise after EEG recording, e.g., by filtering and blind source separation techniques [134, 142, 143].

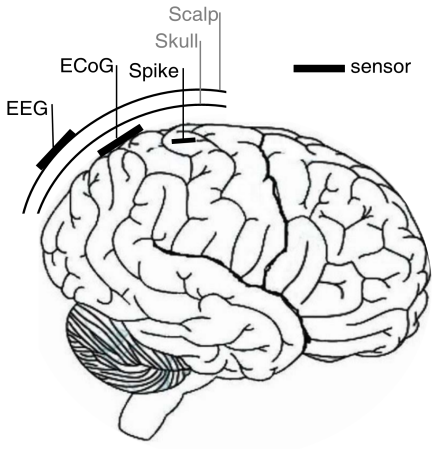


Fig. 1.5: Different methods to measure electrical potentials of brain activity: EEG, ECoG and spike.

Aiming at understanding how the auditory system processes complex auditory stimuli using EEG, a large research effort has focused relating the EEG responses to auditory stimuli, e.g., by estimating the response function of the auditory system using system identification techniques. Traditionally, the response function is estimated by measuring EEG responses at particular times relative to an impulse-like auditory stimulus and averaging over many responses, commonly referred to as the event-related potentials (ERPs) technique [137, 144, 145]. The ERP technique is typically used with the auditory stimuli consisting of discrete syllables or phonemes, which may not be suited to model the neural processing of natural continuous speech, especially when corrupted by noise and reverberation. Therefore, it has recently been proposed to model the auditory system by impulse response functions [146, 147], assuming that the auditory system can be modeled as a linear time-invariant (LTI) system. Although it is unlikely that the complex neural processing in auditory system is performed in a linear and time-invariant fashion, LTI models may still be a reasonable approximation of the linear neural processing [148]. LTI models allow to either predict the EEG responses from a representation of auditory stimulus, referred to as forward modeling [22, 128, 133, 149–152], or to reconstruct a representation of the auditory stimulus from EEG recordings, referred to as backward modeling [16, 22, 128, 133, 153–156]. In the following, we will present an overview of AAD methods using forward and/or backward models.

1.4.2 *Forward-model-based AAD methods*

Forward-model-based methods aim to perform AAD by predicting the measured EEG responses from the envelope of the clean speech of the attended speaker (see Fig. 1.6). Initially, forward models were used to investigate how different representations of the auditory stimulus are encoded in multi-channel EEG responses.

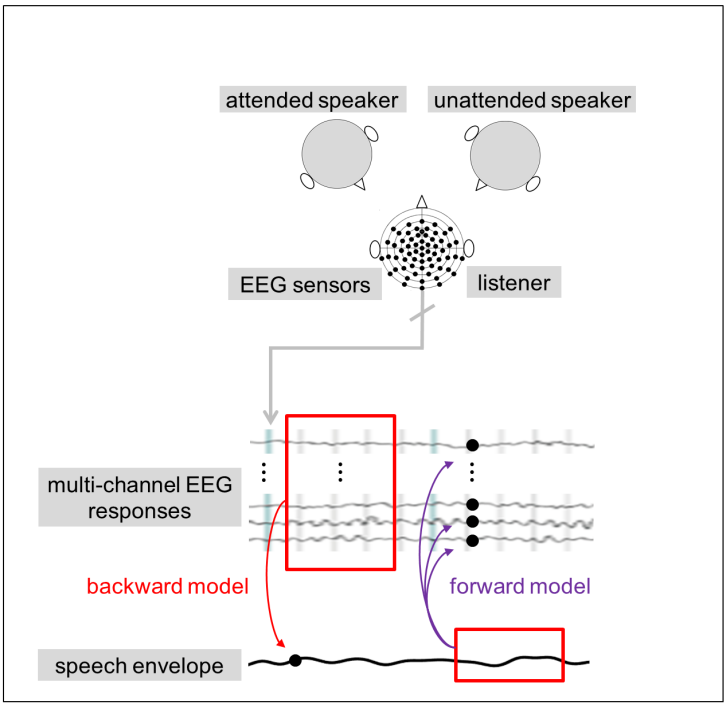


Fig. 1.6: Illustration of the main difference between the forward and the backward model. The exemplary acoustic scenario comprises an attended speaker and an unattended speaker. The ongoing EEG responses of a listener to these acoustic stimuli are recorded.

In [150, 157] it has been proposed to estimate the impulse response function, also known as the temporal response function (TRF), by minimizing the least-squares error between the measured EEG responses and the predicted EEG responses. TRFs relate a representation of the input stimulus to each of the multi-channel EEG responses separately, thereby ignoring information between EEG channels, e.g., inter-channel correlation. Using TRFs, numerous studies have proposed to predict the EEG responses from the slowly varying temporal envelope of a speech signal (< 9 Hz) [150, 152, 158].

A forward-model-based AAD method using TRFs has recently been proposed to identify the attended speaker in an acoustic scenario with two competing speakers [128]. This AAD method aims at predicting each of the multi-channel EEG responses using trained attended TRFs. In the training step, the envelope of the clean speech signal of the attended speaker is used to train the attended TRFs. In the decoding step, the EEG responses are predicted from the envelope of the attended and the unattended speaker using the trained TRFs. These predicted EEG responses are compared with the measured EEG responses by computing correlation coefficients.

Based on these correlation coefficients, the speaker with the largest correlation is identified as the attended speaker. In [22] it has been proposed to combine the correlation coefficients computed for different EEG channels using SVM classifiers to identify the attended speakers.

Since the aforementioned forward-model-based AAD methods are unable to exploit the information between EEG channels to design AAD TRFs, their performance for AAD is typically limited. In addition, since the correlation coefficients using these methods are highly fluctuating, a large correlation window on the order of $30 - 60$ seconds is typically required to achieve a reliable decoding performance. Furthermore, these methods rely on the assumption that the clean speech signals of the attended and the unattended speaker are available as reference signals for decoding.

1.4.3 *Backward-model-based AAD methods*

Backward-model-based methods aim to perform AAD by reconstructing the attended speech envelope from single-trial EEG recordings (see Fig. 1.6). Initially, backward models were mainly used to study the neural processing of auditory scene analysis, e.g., attending to a target speaker in the presence of an interfering speaker [20,133,152,156,158–162]. Backward models reconstruct a representation of the auditory stimulus from multi-channel EEG recordings using a spatio-temporal filter. In [133] it has been proposed to compute the spatio-temporal filter by minimizing the least-squares error between the envelope of the speech stimulus and the reconstructed envelope. The spatio-temporal filter allows to exploit the correlation between the EEG channels for reducing stimulus-irrelevant neural activities that are spatially correlated between channels, e.g., muscle artifacts. In addition, the spatio-temporal filter intrinsically performs EEG channel selection by allocating larger weights to channels highly contributing to the reconstruction and smaller weights to channels with little information related to the reconstruction [156].

Recently, several backward-model-based AAD methods have been proposed based on, e.g., a least-squares cost function [16,18,19,22,28,128,129,163], canonical correlation analysis [21], a state-space model [24] and neural networks [23,130]. The least-squares-based AAD method aims at reconstructing the envelope of the attended speaker from the EEG recordings using a spatio-temporal filter. In the training step, the clean speech signal of the attended speaker is used to train the spatio-temporal filter by minimizing the least-squares error between the attended speech envelope and the reconstructed envelope. In the decoding step, the attended speech envelope is reconstructed from the EEG recordings using the spatio-temporal trained filter. The reconstructed speech envelope is then compared with the envelope of the reference signals by computing correlation coefficients to identify the attended speaker.

To avoid over-fitting, several methods generalizing the least-squares-based AAD methods have been proposed, e.g., by constraining the filter coefficients [16, 18, 19, 22, 128, 129]. In [16, 18, 19, 129] it has been proposed to regularize the least-squares cost function by constraining the squared l_2-norm of either the filter coefficients

or the derivatives of the filter coefficients. While these constraints tend to avoid over-fitting by smoothing the filter coefficients, the filter coefficients that hardly contribute to the envelope reconstruction are still assigned near-to-zero values. To promote sparsity of the filter coefficients, it has hence been proposed in [128] to constrain the l_1-norm of the filter coefficients, i.e., only keep a relatively small numbers of filter coefficients with non-zero values. To control the balance between smoothness and sparsity of the filter coefficients, it has been proposed in [22] to combine the squared l_2-norm and l_1-norm using a trade-off parameter. It has been shown that the regularization with the squared l_2-norm of the derivative of the filter coefficients improves the accuracy of attended envelope reconstruction compared to the regularization with the squared l_1-norm or the squared l_2-norm of the filter coefficients [22]. However, all discussed regularization approaches yield a comparable decoding performance [22].

Using a combination of forward and backward models, another AAD method has been proposed in [21, 128], where a canonical correlation analysis is used to compute an attended and an unattended linear transformation which maximize the mutual projections between the EEG recordings and the attended and the unattended envelope. The few components of the attended and the unattended linear transformation resulting in the largest mutual projections are then used to identify the attended speaker using SVM classifiers. In [128] it has been shown that the canonical-correlation-analysis-based AAD method can typically yield a larger decoding performance compared to the least-squares-based AAD methods.

Aiming at AAD for dynamic scenarios where, e.g., attention switching occurs, it has been proposed to decode AAD based on updating filters [24]. The filters aim to reconstruct both the attended and the unattended speech envelopes from the EEG recordings and are recursively updated with a forgetting factor as the new EEG samples become available. To decode AAD, the l_1-norm of the filter coefficients is translated into a probabilistic measure of the attentional state using a state-space model. Since the parameters of the state-space model are updated using two nested expectation-maximization (EM) algorithms, the proposed AAD method is highly complex.

To take into account the non-linear processing of acoustic signals along the auditory pathway, it has been proposed in [23, 130] to reconstruct the attended speech envelope from the EEG recordings using a non-linear model, more in particular a convolutional DNN. Similarly as for the (linear) least-squares-based AAD method, the DNN-based reconstructed envelope is then correlated with the envelope of the reference signals for decoding. As an alternative to the DNN-based envelope reconstruction, an integrated DNN-based architecture has been proposed in [130]. The integrated DNN-based AAD directly compares the EEG recordings with the envelopes of the reference signals, generating a similarity score for each reference signal. The reference envelope with the largest similarity score is then identified as the attended speech envelope. These DNN-based AAD methods may yield a slightly larger decoding performance compared to the least-squares-based AAD methods. However, generalization of these methods to EEG responses where the listener switches attention between speakers remains to be investigated.

Since backward-modeling-based AAD methods exploit information between the EEG channels, e.g., inter-channel correlation, these methods typically yield a larger decoding performance compared to the forward-model-based AAD methods. Similarly to the forward-model-based AAD methods, most aforementioned backward-model-based AAD methods suffer from highly fluctuating correlation coefficients and therefore rely on a large correlation window on the order of $30-60$ seconds. In addition, most aforementioned backward-modeling-based methods rely on the assumption that the clean speech signals of the attended and the unattended speakers are available as reference signals for decoding.

Challenges and open issues

The performance of most aforementioned AAD methods has been extensively investigated for anechoic acoustic conditions [22,28,128,129,163–165]. However, in practice also background noise and reverberation are present, which may negatively influence the decoding performance. In addition, most aforementioned AAD methods rely on the assumption that the clean speech signals of the speakers are available as reference signals for decoding. However, in hearing aid applications obviously only the microphone signals containing acoustic disturbances (interfering speaker, background noise and reverberation) are available. In order to fully understand the performance of AAD in realistic noisy and reverberant acoustic conditions, one goal of this thesis is to analyze the performance of AAD for different acoustic conditions and to investigate the impact of each disturbance in the microphone signals on AAD performance, which can be used to design efficient algorithms for generating appropriate reference signals for decoding from the microphone signals.

Generally, most aforementioned AAD methods suffer from highly fluctuating correlation coefficients and therefore rely on a large correlation window on the order of $30-60$ seconds. The large correlation window can cause a large processing delay and hence limits the feasibility of AAD for hearing aid applications. Therefore, another goal of this thesis is to improve the decoding performance of the least-squares-based and DNN-based AAD methods using correlation coefficients obtained with a small correlation window.

1.5 Open-loop cognitive-driven speech enhancement using AAD

Aiming at generating appropriate reference signals for decoding from the microphone signals and incorporating AAD in speech enhancement algorithms, several cognitive-driven single- and multi-channel speech enhancement algorithms [27, 28, 30, 166, 167] have been proposed. Most of these cognitive-driven speech enhancement algorithms have been developed for open-loop scenarios, where AAD is performed in an off-line fashion without presenting (on-line) feedback, e.g., the enhanced attended speaker, to the listener. The single-microphone speech enhancement algorithm proposed in [28,167] uses a deep neural network (DNN) to generate reference signals by separating the speakers from the mixture received at the microphone. Using AAD, one of the reference signals is then selected as the enhanced

attended speaker. The cognitive-driven multi-channel Wiener filter (MWF) proposed in [27, 166] generates reference signals from binaural hearing aid microphone signals using multiple multi-channel Wiener filters based on an envelope demixing algorithm. Using EEG recordings and AAD, one of the reference signals is then selected as the enhanced attended speaker. Experimental results in [27, 28, 166–168] show that the proposed cognitive-driven speech enhancement algorithms are able to enhance the attended speaker and strongly suppress the interfering unattended speaker. To directly generate the enhanced attended speaker from binaural hearing aid microphone signals using AAD, it has been proposed in [30] to jointly decode the auditory attention and perform LCMV beamforming. Experimental results in [30] show that the joint AAD and LCMV beamforming is robust to AAD estimation errors and is able to enhance the attended speaker.

Challenges and open issues

While most aforementioned cognitive-driven speech enhancement algorithms are able to strongly suppress the interfering speaker, which is desired to improve speech intelligibility, it may deprive the listener from switching attention. In addition, the binaural MWF changes the spatial impression of the acoustic scene since all sources at the output of the binaural MWF are perceived as coming from the direction of the attended speaker [10, 95], which may lead to a confusion between acoustical and visual information. Therefore, one goal of this thesis is to develop a cognitive-driven binaural beamforming system that enhances the attended speaker and controls the suppression of the unattended speaker while preserving the spatial impression of the acoustic scene.

While (late) reverberation has a negative effect on speech quality and intelligibility, most of the aforementioned cognitive-driven speech enhancement algorithms may not be able to suppress reverberation. Therefore, another goal of this thesis to develop cognitive-driven beamforming system to enhance the attended speaker and jointly suppress the interfering speaker, reverberation and background noise.

The performance of most aforementioned cognitive-driven speech enhancement algorithms has been investigated only for open-loop scenarios with no attention switch between speakers. To investigate the performance of cognitive-driven speech enhancement for hearing aid applications, closing the loop by presenting auditory feedback according to the AAD results in an on-line fashion is very important. Feedback presentation may influence the subsequent intent of the listener and the brain signals that encode that intent. In [163] the feasibility of closed-loop AAD has been demonstrated by presenting the AAD results as visual feedback, either using different colors or a sphere with different radii. However, the feasibility of closed-loop cognitive-driven speech enhancement remains to be investigated. Therefore, another goal of this thesis is to enable the listener to switch attention between the speakers and to interact with a speech enhancement algorithm using a closed-loop cognitive-driven system.

1.6 Outline of the thesis and main contributions

This thesis deals with the problem of identifying and enhancing the target speaker in realistic acoustic environments based on decoding the auditory attention of the listener using single-trial EEG recordings. To this end, we thoroughly analyze the AAD performance in realistic noisy and reverberant environments, we propose novel methods for decoding auditory attention and we propose cognitive-driven speech enhancement algorithms for hearing aid applications.

The main contributions of this thesis are threefold. First, **we analyze the performance of a least-squares-based AAD method by investigating the impact of different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy)**. We show that for all considered acoustic conditions it is possible to decode auditory attention with a considerably large decoding performance but that the decoding performance is significantly affected by the presence background noise and especially the interfering speaker in the reference signals used for decoding. Second, **we propose several open-loop and closed-loop cognitive-driven speech enhancement systems**. The first system is an open-loop cognitive-driven binaural beamformer, aiming at enhancing the target speaker and suppressing the interfering speaker and background noise while preserving the spatial impression of the acoustic scene. In this system a binaural MVDR or LCMV beamformer is cognitively steered based on AAD and the RTFs of the attended and the unattended speaker. The second system is an open-loop cognitive-driven convolutional beamformer, aiming at enhancing the attended speaker and jointly suppressing the interfering speaker, reverberation and background noise. This system combines a neural-network-based mask estimator, wMPDR or wLCMP convolutional beamformers and AAD. The third system is a closed-loop cognitive-driven gain controller, where real-time AAD enables the listener to directly interact with the speech enhancement system. Third, **we propose methods to improve the decoding performance**. More specifically, we propose a reference signal generation approach based on binary masking, which uses binary masks based on directional speech presence probability to discard low-energy intervals which are susceptible to interfering speech and background noise. In addition, we propose an AAD method based on a state-space model, which improves the decoding performance of linear (least-squares-based) and non-linear (DNN-based) methods using small correlation windows.

In the remainder of this section a chapter-by-chapter overview of this thesis is presented, summarizing the main contributions. Additionally, references to the publications that have been produced in the context of this thesis are provided. A structured overview of the thesis is given in Fig. 1.7.

In **Chapter 2**, we investigate the performance of a least-squares-based AAD method for different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy). In particular, we investigate the impact of different acoustic conditions for AAD filter training and decoding. In addition, we investigate the influence on the decoding performance of the head shadow effect and of the different acoustic disturbances (interfering speaker, noise and reverberation) in the reference

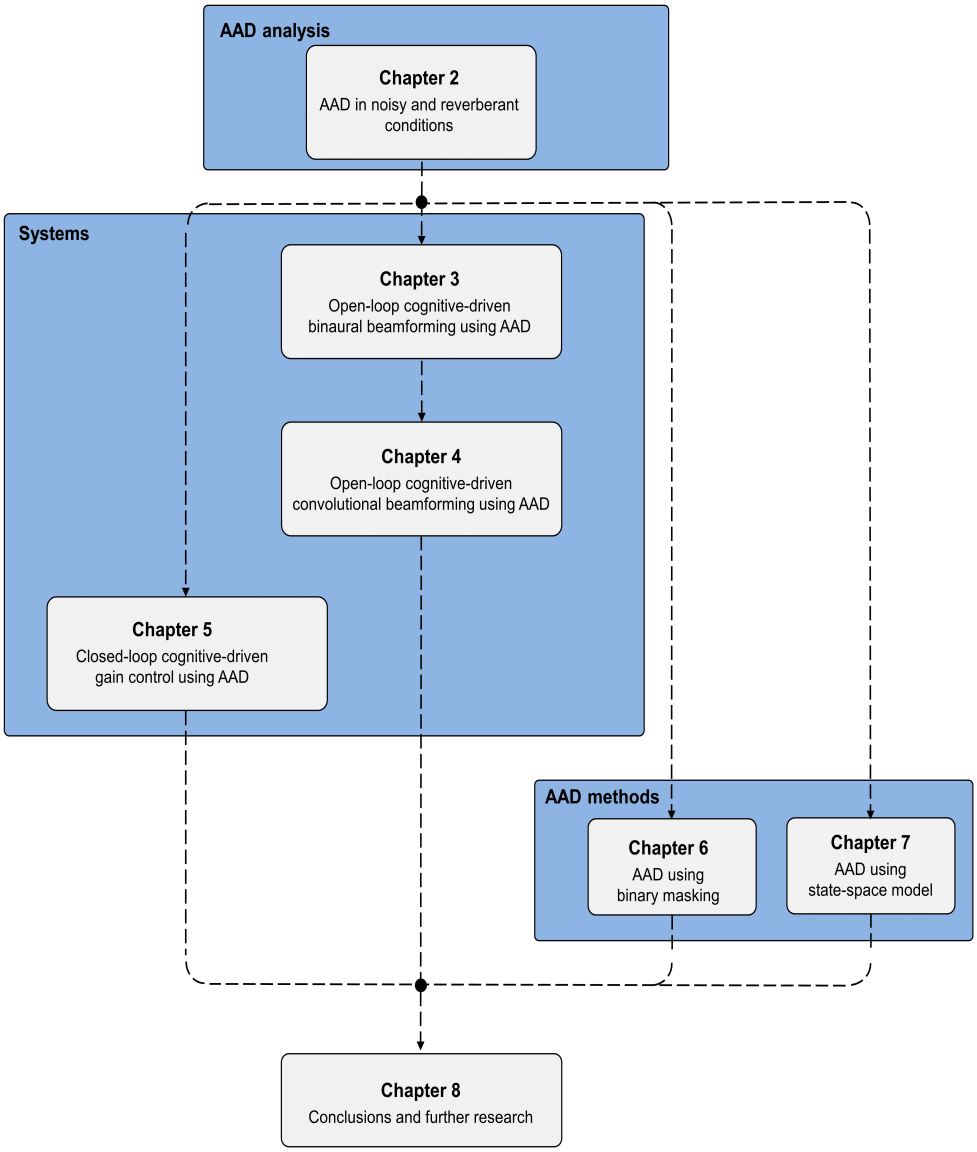


Fig. 1.7: Structure of the thesis.

signals used for decoding and the training signals used for computing the filters. To investigate the performance of the least-squares-based AAD method, we use correlation coefficients obtained with a 60-second correlation window. We show that for all considered acoustic conditions it is possible to decode auditory attention with a considerably large decoding performance ($> 90\%$), even when the acoustic conditions for AAD filter training and decoding are different. We show that for all considered acoustic conditions the head shadow effect has no significant impact on the decoding performance. In addition, we show that the decoding performance ($> 87\%$) is significantly affected by the presence of background noise and especially the interfering speaker in the reference signals used for decoding. Furthermore, we show that it is even feasible to use training signals affected by reverberation, background noise and/or the interfering speaker for computing the filters. Even when using the microphone signals as training signal and reference signals, it is still feasible to perform AAD with a large decoding performance ($> 82\%$). This chapter has been published as a journal paper in the IEEE Transactions on Neural Systems and Rehabilitation Engineering [26]. Other publications related to this chapter are [32, 169].

In **Chapter 3**, we propose an open-loop cognitive-driven binaural speech enhancement system, aiming at enhancing the attended speaker and controlling the suppression of the unattended speaker while preserving the spatial impression of the acoustic scene. First, either the anechoic or the reverberant RTFs of both speakers are estimated from the microphone signals, where the anechoic RTFs are computed based on the estimated DOAs of both speakers. Based on these RTFs, two MVDR or LCMV beamformers are used to generate reference signals for auditory attention decoding. Using the envelopes of these reference signals, the EEG recordings and a 30-second correlation window, in the AAD step the attended and the unattended speaker are identified based on the least-squares-based AAD method, enabling to steer a binaural MVDR or LCMV beamformer. We provide a detailed analysis and experimental comparison between the cognitive-driven binaural LCMV and MVDR beamformers for an acoustic scenario comprising two competing speakers and diffuse background noise in an anechoic and a reverberant condition. In addition, we investigate the impact of RTF and DOA estimation errors and AAD errors on the speech enhancement performance. We show that the proposed system using anechoic DOA-based RTFs significantly improves the binaural SINR for the anechoic condition (7.4–8.7 dB) as well as for the reverberant condition (3.2–3.6 dB) compared to a fixed forward-steered binaural MVDR beamformer (0.3 dB for the anechoic condition and 0.5 dB for the reverberant condition). In particular, the cognitive-driven binaural LCMV beamformer is able to both improve the binaural SINR as well as preserve the binaural cues of both the attended and the unattended speaker. We show that the proposed system using estimated DOA-based anechoic RTFs yields a larger binaural SINR improvement and AAD performance for the reverberant condition compared to using estimated reverberant RTFs. The decoding performance for the LCMV beamformers ($> 82\%$) is larger than for the MVDR beamformers ($> 77\%$). Moreover, for the considered experimental setup we show that the AAD performance and the binaural SINR improvement of the proposed system are sensitive to RTF estimation errors and AAD errors, but not to DOA estimation errors. This chapter has been published as a journal paper in the IEEE Transactions on

Audio, Speech, and Language Processing [31]. Other publications related to this chapter are [29, 170].

While the open-loop cognitive-driven convolutional beamforming system in Chapter 3 is able to suppress the interfering speaker and background noise, it is not designed to suppress reverberation. In **Chapter 4**, we hence propose an open-loop cognitive-driven convolutional beamforming system. Compared to the cognitive-driven binaural beamforming system in Chapter 3, the proposed cognitive-driven convolutional beamforming system allows to enhance the attended speaker and jointly suppress reverberation, the interfering speaker and background noise. In addition, instead of estimating the anechoic or reverberant RTFs based on DOAs, the proposed system estimates the RTFs based on masks. First, the masks of both speakers are estimated from the noisy and reverberant microphone signals using a speech separation neural network. Based on the estimated masks, two convolutional beamformers generate reference signals for AAD by enhancing the speech signal of each speaker. Using a 30-second correlation window, the least-squares-based AAD method then selects one of the reference signals as the enhanced attended speech signal. For the beamformers we propose to use a wMPDR convolutional beamformer as it combines dereverberation, noise suppression and interfering speaker suppression. We also propose an extension of the wMPDR convolutional beamformer, referred to as wLCMP convolutional beamformer, which allows to control the level of suppression of the interfering speaker. We experimentally compare the proposed cognitive-driven convolutional beamforming system with a cognitive-driven speech enhancement systems based on (conventional) MVDR, LCMV, MPDR and LCMP beamformers. We show that the wMPDR and wLCMP convolutional beamformers yield the highest frequency-weighted segmental SNR (fwSSNR) improvement for the anechoic condition as well as for the reverberant condition. For the reverberant condition, only the proposed system using convolutional beamformers provides a fwSSNR improvement, showing the influence of dereverberation, while the system using MVDR, LCMV, MPDR or LCMP beamformers even tend to degrade the fwSSNR. This chapter has been published as a conference paper to the IEEE International Workshop on Machine Learning for Signal Processing (MLSP) [171].

In contrast to Chapters 3 and 4 where we consider open-loop cognitive-driven speech enhancement systems, in **Chapter 5** we propose a closed-loop gain controller system which cognitively steers an adaptive gain controller (AGC) based on real-time AAD for a scenario with two competing speakers. Based on the EEG responses and the (assumed to be known) speech signals of both speakers, the correlation coefficients are generated either using a small correlation window of length 0.25 seconds or using a large correlation window of length 15 seconds. These correlation coefficients are translated into more reliable probabilistic attention measures, based on which the attended and the unattended speaker are identified. The AGC then amplifies the identified attended speaker and attenuates the identified unattended speaker and presents these signals via loudspeakers. To translate the correlation coefficients into probabilistic attention measures, we propose an AAD algorithm using either a generalized linear model (GLM) or a state-space model (SSM). Experimental results demonstrate the feasibility of the proposed closed-loop cognitive-driven gain controller system (both using GLM and SSM), enabling the listener to inter-

act with the system in real-time. Although there is a significant delay to detect attention switches (11.5–19.8 seconds), which causes the attended speaker to be wrongly attenuated and the unattended speaker to be wrongly amplified, the proposed closed-loop system is able to improve the SIR between the attended and the unattended speaker (0.6–3.8 dB). This may make it easier to follow the attended speaker, ignore the unattended speaker and switch attention between both speakers, resulting in a lower cognitive effort compared to open-loop AAD. A statistical analysis of the results show that there is no significant difference in terms of decoding performance, switch detection delay and perceived level of cognitive effort between the open-loop and the closed-loop AAD system. This chapter has been submitted as a journal paper to the Journal of Neural Engineering.

In **Chapter 6**, we propose a novel reference signal generation approach for AAD, which uses binary masks to discard low-energy intervals which are susceptible to interfering speech and background noise. In contrast to Chapters 3, 4 and 5 where we consider complete cognitive-driven speech enhancement systems, in Chapter 6 we focus only on reference signal generation for AAD. First, the directional speech presence probability (DSPP) and the DOAs of both speakers are estimated from the microphone signals. Based on the estimated DSPPs, the intervals with low estimated speech energy for both speakers are detected, resulting in binary masks. The reference signals for decoding are then generated by either masking the microphone signals or the MVDR output signals. In addition, the estimated binary masks are considered themselves as reference signals for decoding. To investigate the performance of the proposed reference signal generation approach, we use correlation coefficients obtained with a 30-second correlation window. We show that the proposed reference signal generation approach significantly improves the decoding performance ($> 81\%$) compared to using the (non-masked) microphone signals ($> 77\%$) and the MVDR output signals ($> 78\%$), especially in the reverberant condition. Quite remarkably, we also show that using the binary masks as reference signals yields a comparable decoding performance to using the masked MVDR output signals. This chapter will be submitted as a journal paper to the IEEE Transactions on Neural Systems and Rehabilitation Engineering.

In **Chapter 7**, we propose a novel AAD method based on a state-space model, which improves the decoding performance of a linear (least-squares-based) AAD method and a non-linear (DNN-based) AAD method using a 5-second correlation window. The state-space model translates the generated correlation coefficients into more reliable probabilistic attention measures, based on which the attended speaker is identified. We experimentally compare the performance of the linear and non-linear AAD methods with and without the state-space model for different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy). We show that without the state-space model a relatively low decoding performance (69%–73%) is obtained, mainly due to the relatively small (attended and unattended) correlation coefficients with a large variability. When using the state-space model, the decoding performance significantly increases, where the increase is considerably larger for the least-squares-based AAD method ($> 94\%$) than for the DNN-based AAD method ($> 73\%$). This chapter has been published as a conference paper in the Proceedings

of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) [172].

In **Chapter 8**, we summarize the main contributions of the thesis and suggest possible topics for further research.

2

IMPACT OF DIFFERENT ACOUSTIC COMPONENTS ON EEG-BASED AUDITORY ATTENTION DECODING IN NOISY AND REVERBERANT CONDITIONS

Identifying the target speaker in hearing aid applications is an essential ingredient to improve speech intelligibility. Recently, a least-squares-based method has been proposed to identify the attended speaker from single-trial EEG recordings for an acoustic scenario with two competing speakers. This least-squares-based auditory attention decoding (AAD) method aims at decoding auditory attention by reconstructing the attended speech envelope from the EEG recordings using a trained spatio-temporal filter. While the performance of this AAD method has been mainly studied for noiseless and anechoic acoustic conditions, it is important to fully understand its performance in realistic noisy and reverberant acoustic conditions. In this paper, we investigate AAD using EEG recordings for different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy). In particular, we investigate the impact of different acoustic conditions for AAD filter training and for decoding. In addition, we investigate the influence on the decoding performance of the different acoustic components (i.e. reverberation, background noise and interfering speaker) in the reference signals used for decoding and the training signals used for computing the filters. First, we found that for all considered acoustic conditions it is possible to decode auditory attention with a considerably large decoding performance. In particular, even when the acoustic conditions for AAD filter training and for decoding are different, the decoding performance is still comparably large. Second, when using speech signals affected by either reverberation and/or background noise there is no significant difference in decoding performance ($p > 0.05$) compared to when using clean speech signals as reference signals. In contrast, when using reference signals affected by the interfering speaker, the decoding performance significantly decreases. Third, the experimental results indicate that it is even feasible to use training signals affected by reverberation, background noise and/or the interfering speaker for computing the filters.

2.1 Introduction

In complex acoustic conditions the human auditory system has a remarkable ability to segregate a speaker of interest from a mixture of speakers and background noise [45, 173]. In contrast with normal-hearing persons, hearing-impaired persons typically have more difficulties with such auditory segregation, particularly in multi-talker scenarios [174]. Although many acoustic signal processing algorithms are available to reduce background noise or to perform source separation in multi-talker scenarios [10, 13], these algorithms typically need to rely on assumptions about the target speaker to be enhanced. For example, in hearing aid applications the target speaker is typically assumed to be located in front of the user or is assumed to be the loudest speaker. As in real-world conditions such assumptions are often violated, the performance of these algorithms may substantially decrease. Therefore, successfully identifying the target speaker in hearing aid applications is very important to improve speech intelligibility.

Recent studies have shown that auditory cortical responses are correlated with the envelope of the attended speech signal [152, 175, 176], based on which decoding and encoding properties of the speech signal, e.g. spectrotemporal features and perceptual unites, have been studied in the brain auditory pathway [177,178]. Based on this finding, an auditory attention decoding (AAD) method has been proposed in [16] to identify the attended speaker from single-trial EEG recordings. This method aims at reconstructing the attended speech envelope from the EEG recordings using a trained spatio-temporal filter. In the *training step*, the clean speech signal of the attended speaker is used to train a spatio-temporal filter by minimizing the least-squares error between the attended speech envelope and the reconstructed envelope. In the *decoding step*, the clean speech signals of both the attended and the unattended speaker are used as reference signals. In [16] it has been shown that for high-density EEG recordings it is possible to decode auditory attention when presenting the clean speech signals of the different speakers to different ears of a listener (i.e. dichotic stimuli presentation). When presenting competing speech signals in a simulated anechoic condition including head filtering effects, it has been shown in [164] that a larger AAD performance can be obtained compared to dichotic presentation. Recently, a large research effort has focused on investigating how to use AAD as part of a brain-computer interface for real-world applications, e.g., to control a hearing aid [18, 19, 21, 25, 27–29, 164, 165, 179–182], mainly however for anechoic conditions. Aiming at integrating a small-size EEG recording system in hearing aids, in [18, 165, 179] the reliability of AAD using a low number of EEG electrodes has been shown in an anechoic condition. Aiming at investigating the effect of neurofeedback, in [180] the feasibility of an online closed-loop system for AAD has been shown in an anechoic condition. Instead of using the clean speech signals of the attended and the unattended speaker as reference signals for decoding, in [19, 27–29, 181] the effect of different reference signals on the AAD performance has been investigated for an anechoic condition. Using simulated noisy reference signals for decoding, in [19] we have investigated the robustness of AAD to residual interference and background noise. In [27, 181] a neuro-steered noise reduction algorithm has been proposed to suppress the unattended speaker based on the AAD

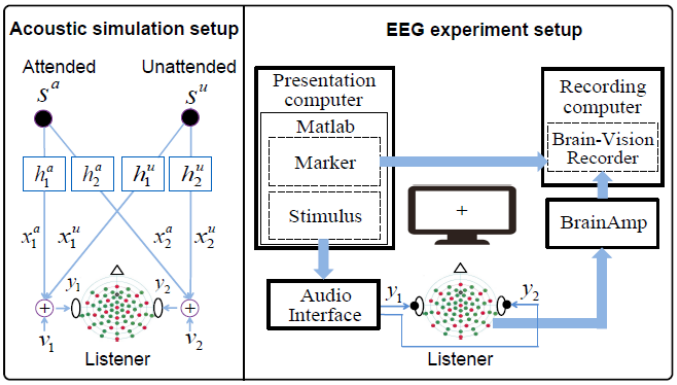


Fig. 2.1: Acoustic simulation setup and EEG experiment setup. The acoustic simulation setup was used for simulating the presented stimuli in different acoustic conditions. For the EEG experiment setup, MATLAB was used for sending the acoustic stimuli to the audio interface and the event markers to the Brain-Vison recorder software. The acoustic stimuli were presented to the participants via earphones using the audio interface. The EEG responses were amplified using BrainAmp and recorded together with the event markers using Brain-Vision.

decision for an anechoic condition. In [28] an AAD-based sound source separation algorithm using deep neural networks has been presented to suppress the unattended speaker. In [29] we have investigated steerable beamformers to generate reference signals for AAD in an anechoic condition.

While the performance of the aforementioned least-squares-based AAD method has been extensively investigated for noiseless and anechoic acoustic conditions, in practice also background noise and reverberation, i.e. acoustic reflections against walls and objects, are present. Reverberation is known to spectro-temporally distort speech signals, causing the binaural spatial cues and pitch to become less reliable for performing auditory attention tasks [183–186]. In addition, interfering speakers and background noise degrade the attended speech signal, possibly leading to a severe speech encoding degradation at the level of the auditory nerve and the brainstem [160, 187]. Since in noisy and reverberant conditions the available signals at the ears contain several acoustic components (i.e. reverberation, background noise and interfering speaker), fully understanding the impact of each acoustic component on AAD is of crucial importance, e.g., in order to generate appropriate reference signals for decoding from these signals. Recently, in [20] the performance of the least-squares-based AAD method was investigated for noisy and reverberant acoustic conditions. In [20] the same acoustic condition was used for AAD filter training and for decoding and the feasibility of using reverberant speech signals both as training and as reference signals was investigated. It was shown that in this way a comparable decoding performance for the reverberant condition as for the anechoic condition can be obtained. In this paper, we perform a more detailed analysis of the performance of the least-squares-based AAD method for an acoustic scenario comprising two competing speakers, background noise and reverberation. Compared

Table 2.1: Acoustic signals used for experimental analysis

Signal	Definition
s^a, s^u	clean speech signal
$x_m^{a,an}$, $x_m^{u,an}$	anechoic speech signal
x_m^a, x_m^u	reverberant speech signal
x_m^{an}	interfered speech signal
$x_m^{a,no}$, $x_m^{u,no}$	noisy speech signal
y_m	binaural speech signal

to [20] we consider more acoustic conditions, especially with regard to background noise, and we specifically investigate the impact of different acoustic conditions for the training and the decoding steps. In addition, we investigate the influence on the decoding performance of the different acoustic components in the reference signals used for decoding and the training signals used for computing the filters. Some preliminary results were presented in [169], where we investigated the feasibility of using the (unprocessed) signals at the ears, containing reverberation, background noise and the interfering speaker, as reference and training signals.

The paper is organized as follows. In Section 2.2 the different acoustic conditions used for recording the EEG responses and the different acoustic signals used for the experimental analysis are introduced. In Section 2.3 the training and decoding steps of the least-squares-based AAD method are briefly reviewed. Section 2.4 describes the acoustic and EEG measurement setup used for the experiments. In Section 2.5 the experimental results are presented and discussed, exploring the influence on the decoding performance of the different acoustic conditions and acoustic components.

2.2 Acoustic Conditions and Components

We consider an acoustic scenario comprising two competing speakers and background noise in a reverberant environment (see left part of Fig. 2.1). The clean speech signal of the attended speaker is denoted as $s^a[i]$, while the clean speech signal of the unattended speaker is denoted as $s^u[i]$, with i the discrete time index. The signals at the ears of the listener consist of a mixture of both speakers, including head filtering effects, reverberation and background noise. The signal $y_m[i]$ at the m-th ear, with $m = 1$ denoting the left ear and $m = 2$ denoting the right ear, can be written as

$$y_m[i] = \underbrace{h_m^a[i] * s^a[i]}_{x_m^a[i]} + \underbrace{h_m^u[i] * s^u[i]}_{x_m^u[i]} + v_m[i], \tag{2.1}$$

where $h_m^a[i]$ and $h_m^u[i]$ denote the (reverberant) acoustic impulse response between the m-th ear and the attended and the unattended speaker, respectively, $*$ denotes

the convolution operation, and $v_m[i]$ denotes the background noise component at the m-th ear. The reverberant speech signal of the attended and the unattended speaker at the m-th ear is denoted as $x_m^a[i]$ and $x_m^u[i]$, respectively. These reverberant speech signals consist of an anechoic speech signal encompassing the (anechoic) head filtering effect, i.e. $x_m^{a,an}[i]$ and $x_m^{u,an}[i]$, and a reverberation component. For notational conciseness the index i will be omitted in the remainder of this paper, except where explicitly required.

For the EEG recordings we will consider four different acoustic conditions, i.e. anechoic, reverberant, noisy and reverberant-noisy. We refer to the EEG data recorded in a specific acoustic condition as the *EEG condition*. Depending on the acoustic condition, the stimuli presented at the ears of the listener obviously comprise different acoustic components:

- in the *anechoic* condition (an), the mixture of the anechoic speech signals of the attended and the unattended speaker is presented.
- in the *noisy* condition (no), the mixture of the anechoic speech signals of the attended and the unattended speaker and background noise is presented.
- in the *reverberant* condition (re), the mixture of the reverberant speech signals of the attended and the unattended speaker is presented.
- in the *reverberant-noisy* condition (rn), the mixture of the reverberant speech signals of the attended and the unattended speaker and background noise is presented.

To investigate the impact of the different acoustic components on the AAD performance, we will consider several acoustic signals (see Table 2.1) to compute envelopes for filter training and evaluation:

- the clean speech signals s^a and s^u.
- the anechoic speech signals $x_m^{a,an}$ and $x_m^{u,an}$, i.e. the clean speech signals affected by head filtering effects.
- the reverberant speech signals x_m^a and x_m^u, i.e. the anechoic speech signals affected by reverberation.
- the interfered speech signals, i.e. the anechoic speech signals affected by an interfering speaker
$$x_m^{an} = x_m^{a,an} + x_m^{u,an}. \tag{2.2}$$
- the noisy speech signals, i.e. the anechoic speech signals affected by background noise
$$x_m^{a,no} = x_m^{a,an} + v_m, \quad x_m^{u,no} = x_m^{u,an} + v_m. \tag{2.3}$$
- the binaural speech signals y_m in (2.1), i.e. the anechoic speech signals affected by reverberation, background noise and an interfering speaker.

It should be noted that in the experiments (see Section 2.4.2) the positions of the attended and the unattended speaker are not always the same, i.e. for some participants the attended speaker is on the right side (and the unattended speaker on the left side), whereas for some participants the attended speaker is on the left side (and the unattended speaker on the right side). Due to the head filtering effect, the broadband energy ratio between the attended speech component and

the unattended speech component in the signals at the ears is always smaller at the side of the unattended speaker than at the side of the attended speaker for the considered scenario. Therefore, the speech signals in Table 2.1 at the side of the attended speaker will be referred to as attended speech signals and the speech signals at the side of the unattended speaker as unattended speech signals.

2.3 Auditory Attention Decoding Method

This section briefly reviews the least-squares-based AAD method proposed in [16]. This method aims at reconstructing the attended speech envelope from the EEG recordings using a trained spatio-temporal filter. Section 2.3.1 describes the training step, where the envelope of a training signal is used together with the EEG recordings to compute the filter. Section 2.3.2 describes the decoding step, where the envelopes of two reference signals (attended and unattended) are compared with an estimate of the attended speech envelope computed using the trained filter.

2.3.1 *Training Step*

In the training step, the attended speaker is assumed to be known and an attended speech signal (e.g., the clean speech signal of the attended speaker s^a) is used as training signal. From this signal the attended speech envelope $e^a[k]$, with $k = 1 \dots K$ the sub-sampled time index, is extracted, e.g., based on the Hilbert transform [129]. The attended speech envelope is then estimated from the EEG recordings $r_c[k]$, $c = 1 \dots C$, using a spatio-temporal filter as

$$\hat{e}^a[k] = \sum_{c=1}^{C} \sum_{l=0}^{L-1} g_{c,l}\ r_c[k + l + \Delta], \tag{2.4}$$

with $g_{c,l}$ the l-th filter coefficient in the c-th channel, L the number of filter coefficients per channel, and Δ modeling the latency of the attentional effect in the EEG responses to the speech stimuli. In vector notation, (2.4) can be written as

$$\hat{e}^a[k] = \mathbf{g}^T \mathbf{r}[k], \tag{2.5}$$

with

$$\mathbf{g} = \left[\mathbf{g}_1^T\ \mathbf{g}_2^T \dots\ \mathbf{g}_C^T\right]^T, \tag{2.6}$$

$$\mathbf{g}_c = \left[g_{c,0}\ \ g_{c,1} \cdots\ g_{c,L-1}\right]^T, \tag{2.7}$$

$$\mathbf{r}[k] = \left[\mathbf{r}_1^T[k]\ \mathbf{r}_2^T[k]\ \dots\ \mathbf{r}_C^T[k]\right]^T, \tag{2.8}$$

$$\mathbf{r}_c[k] = \left[r_c[k + \Delta]\ r_c[k + 1 + \Delta]\ \dots\ r_c[k + L - 1 + \Delta]\right]^T, \tag{2.9}$$

with $(.)^T$ denoting the transpose operation. The spatio-temporal filter $\mathbf{g}$ is computed by minimizing the least-squares error between the attended speech envelope $e^a[k]$

and the reconstructed envelope $\hat{e}^a[k]$, regularized with the squared l_2-norm of the derivatives of the filter coefficients to avoid over-fitting [16, 18, 129, 169], i.e.

$$J(\mathbf{g}) = \frac{1}{K}\sum_{k=1}^{K}\left(e^a[k] - \mathbf{g}^T\mathbf{r}[k]\right)^2 + \beta\mathbf{g}^T\mathbf{D}\mathbf{g}, \tag{2.10}$$

with $\mathbf{D}$ denoting the derivative matrix [18] and β denoting a regularization parameter. The filter minimizing the regularized least-squares cost function in (2.10) is equal to

$$\mathbf{g} = (\mathbf{Q} + \beta\mathbf{D})^{-1}\mathbf{q}, \tag{2.11}$$

with the correlation matrix $\mathbf{Q}$ and the cross-correlation vector $\mathbf{q}$ given by

$$\mathbf{Q} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{r}[k]\,\mathbf{r}^T[k]\right), \quad \mathbf{q} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{r}[k]\,e^a[k]\right). \tag{2.12}$$

In this paper we will consider several *EEG training conditions* (tc) for computing the filter $\mathbf{g}$, i.e. $tc = an$ using EEG responses recorded in the anechoic condition, $tc = re$ using EEG responses recorded in the reverberant condition, $tc = no$ using EEG responses recorded in the noisy condition, and $tc = rn$ using EEG responses recorded in the reverberant-noisy condition. In addition, we will consider the EEG training condition $tc = ac$, in which EEG responses from all conditions are used for computing the filter.

Aiming at investigating the influence of each acoustic component, in this paper we will consider different attended speech signals (see Table 2.1) as *training signals*, more in particular the clean attended speech signal s^a, the anechoic attended speech signal $x_m^{a,an}$, the reverberant attended speech signal x_m^a, the interfered attended speech signal x_m^{an}, the noisy attended speech signal $x_m^{a,no}$, and the binaural attended speech signal y_m.

2.3.2 *Decoding Step*

For each acoustic condition, the complete set of EEG responses is segmented into T trials (see Section 2.4.4 for more details). To decode to which speaker a listener attended during trial t, first an estimate of the attended speech envelope $\hat{e}_t^a[k]$ is computed using the (trained) filter $\mathbf{g}_t$, i.e.

$$\hat{e}_t^a[k] = (\mathbf{g}_t)^T\mathbf{r}_t[k], \tag{2.13}$$

with $\mathbf{r}_t[k]$ denoting the EEG recordings of trial t. Next, the correlation coefficients between the estimated attended speech envelope $\hat{e}_t^a[k]$ and the envelope of two reference signals, i.e. namely the attended and the unattended reference signal, are computed as

$$\rho_t^a = \rho(e_t^a[k],\, \hat{e}_t^a[k]), \quad \rho_t^u = \rho(e_t^u[k],\, \hat{e}_t^a[k]), \tag{2.14}$$

where ρ_t^a and ρ_t^u denote the attended and the unattended correlation coefficient, respectively, and $e_t^a[k]$ and $e_t^u[k]$ denote the attended and the unattended speech envelope, respectively. When $\rho_t^a > \rho_t^u$, it is decided that auditory attention has been correctly decoded. Accordingly, a larger difference between the attended and the unattended correlation coefficient $\rho_t^a - \rho_t^u$ (referred to as correlation difference) is indicative of a more reliable AAD decision. The decoding performance P is defined as the percentage of correctly decoded trials over all considered trials and all participants. To compute the correlation coefficients in (2.14), EEG recordings in different acoustic conditions can be used for computing $\hat{e}_t^a[k]$. In addition, aiming at investigating the influence of each acoustic component on the decoding performance, different reference signals (see Table 2.1) can be used for computing the attended and the unattended speech envelope $e_t^a[k]$ and $e_t^u[k]$, respectively.

In this paper we will investigate the decoding performance for several *EEG evaluation conditions* $ec \in \{an, re, no, rn, ac\}$, with P_{ec} denoting the decoding performance for a specific EEG evaluation condition. To decode trial t of an EEG evaluation condition using the filter trained in a specific EEG training condition which is not necessary the same as the EEG evaluation condition, the filter $\mathbf{g}_t$ is computed as follows:

- when the trial t to be decoded is part of the trials in the EEG training condition, the filter is computed using (2.11) as

$$\mathbf{g}_t = \left(\tilde{\mathbf{Q}}_t + \beta\mathbf{D}\right)^{-1}\tilde{\mathbf{q}}_t, \tag{2.15}$$

  with $\tilde{\mathbf{Q}}_t$ the average correlation matrix, computed by averaging all correlation matrices corresponding to trials in the EEG training condition *except* trial t, and $\tilde{\mathbf{q}}_t$ the average cross-correlation vector, computed by averaging all cross-correlation vectors corresponding to trials in the EEG training condition *except* trial t, i.e.

$$\tilde{\mathbf{Q}}_t = \frac{1}{T-1}\sum_{n=1,\, n\neq t}^{T}\mathbf{Q}_n, \quad \tilde{\mathbf{q}}_t = \frac{1}{T-1}\sum_{n=1,\, n\neq t}^{T}\mathbf{q}_n. \tag{2.16}$$

  This procedure corresponds to leave-one-out cross validation.

- when the trial t to be decoded is not part of the trials in the EEG training condition, the filter is computed using (2.11) as

$$\mathbf{g}_t = \left(\bar{\mathbf{Q}} + \beta\mathbf{D}\right)^{-1}\bar{\mathbf{q}}, \tag{2.17}$$

  with $\bar{\mathbf{Q}}$ the average correlation matrix, computed by averaging all correlation matrices corresponding to trials in the EEG training condition, and $\bar{\mathbf{q}}$ the average cross-correlation vector, computed by averaging all cross-correlation vectors corresponding to trials in the EEG training condition, i.e.

$$\bar{\mathbf{Q}} = \frac{1}{T}\sum_{n=1}^{T}\mathbf{Q}_n, \quad \bar{\mathbf{q}} = \frac{1}{T}\sum_{n=1}^{T}\mathbf{q}_n, \tag{2.18}$$

Since the number of trials across acoustic conditions is different (see Section 2.4.2), for $tc = ac$ the average correlation matrix and the average cross-correlation vector ($\tilde{\mathbf{Q}}_t$, $\bar{\mathbf{Q}}$, $\tilde{\mathbf{q}}_t$ and $\bar{\mathbf{q}}$) are computed in such a way that the contribution of trials from each acoustic condition is considered equally.

In [19] it has been shown that the parameters involved in the filter design (Δ, L, β) play an important role in obtaining a good decoding performance. In order not to favour one specific EEG evaluation condition, the filter parameters have been determined to optimize the average decoding performance P_{ac} over all considered acoustic conditions. Please note that the filter parameters have been optimized per participant and for each EEG training condition (see Section 2.4.2 and 2.4.4).

2.4 Acoustic and EEG Measurement Setup

In this section, we describe the acoustic and EEG measurement setup used for the experiments and information about the participants and the used paradigm.

2.4.1 *Participants*

Eighteen native German-speaking participants (right-handed and aged between 21 and 34 years) took part in this study. All participants were normal-hearing as was confirmed by pure tone audiometry. The participants reported no past or present neurological or psychiatric conditions. All participants signed an informed consent form and were paid for their participation. Two participants were excluded from the analysis, one participant due to poor attentional performance (as revealed by the questionnaire results) and the other participant due to a technical hardware problem.

2.4.2 *Acoustic Stimuli*

Two German audio stories, uttered by two different male speakers, were used as the clean speech signals (sampling frequency of 16 kHz). One story was from the German audio book website [188] and the other story was from a selection of audio books [189]. Speech pauses that exceeded 0.5 s were shortened to 0.5 s. Before performing the experiment, the participants reported no, or very limited, knowledge of the audio stories. The acoustic stimuli were simulated by convolving the clean speech signals (i.e. the audio stories) with non-individualized binaural acoustic impulse responses, either from [110, 190], or [191], and by adding diffuse babble noise, generated according to [192]. The competing speakers were simulated at $-45°$ (left) and $45°$ (right). Eight different acoustic conditions were considered for the stimuli (see Table 2.2): anechoic, reverberant with a moderate and a large reverberation time ($T_{60} = 0.5$ s, $T_{60} = 1$ s), noisy with two different broadband signal-to-noise ratios (SNR = 9.0 dB, SNR = 4.0 dB), and three combinations of reverberation and noise. The SNR is defined as the broadband energy ratio between the reverber-

Table 2.2: Acoustic conditions used for experimental analysis and stimuli presentation

Experimental Analysis Condition	**Stimuli Presentation**	**SNR** [dB]	$\boldsymbol{T_{60}}$ [s]	**Number of Trials**
Anechoic (*an*)	Anechoic [110]	∞	< 0.05	40
Reverberant (*re*)	Reverberant I [110]	∞	0.50	10
	Reverberant II [190, 191]	∞	1.00	10
Noisy (*no*)	Noisy I [110]	9.0	< 0.05	10
	Noisy II [110]	4.0	< 0.05	10
Reverberant-noisy (*rn*)	Reverberant-noisy I [110]	9.0	0.50	10
	Reverberant-noisy II [110]	4.0	0.50	10
	Reverberant-noisy III [190, 191]	9.0	1.00	10

ant speech signal of the attended and the unattended speaker at the ears and the background noise component at the ears, i.e.

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_i |x_1^a[i]|^2 + |x_1^u[i]|^2 + |x_2^a[i]|^2 + |x_2^u[i]|^2}{\sum_i |v_1[i]|^2 + |v_2[i]|^2}. \tag{2.19}$$

For the experimental analysis, the acoustic conditions were grouped based on acoustic similarity as shown in Table 2.2, resulting in four experimental analysis conditions, i.e. anechoic, reverberant, noisy, and reverberant-noisy. The acoustic stimuli were presented to the participants via insert earphones (E-A-RTONE 3A) using an RME HDSP 9632 PCI Audio Interface, Tucker Davis Technologies programmable attenuators, and MATLAB, which was also used for generating the EEG marker stream (see Fig. 2.1).

2.4.3 *Paradigm*

The stimuli were presented in 11 sessions, each of length 10 minutes, interrupted by short breaks. Among all participants, 8 participants were instructed to attend to the left speaker, while 10 participants were instructed to attend to the right speaker. The participants were also instructed to look at a fixation cross on a screen and minimize eye blinking. For each participant, the anechoic condition was always assigned to the first session and subsequently to every other third session (i.e. session 4, 7, and 10). Aiming at minimizing the influence of the speech material on AAD, the acoustic conditions (except for the anechoic condition) were randomly assigned to the other sessions. Following each session, the participants were asked to fill out a questionnaire consisting of 10 multiple-choice questions related to each story. The questionnaire was aimed to indicate whether the participants attended to the instructed speaker and whether the audio story was intelligible in the different acoustic conditions. The experiment for each participant took place on two different days.

2.4.4 *EEG Setup and Signal Pre-processing*

The EEG responses were recorded using a BrainAmp system, provided by Brain-Products GmbH, Germany, and $C = 64$ channels, provided by Easycap GmbH, Germany, with a sampling frequency of 500 Hz (see EEG experiment setup in Fig. 2.1). The EEG responses were referenced to the nose electrode and recorded using the Brain-Vision recorder software. The EEG recordings were re-referenced offline to a common average reference, band-pass filtered between 2 Hz and 8 Hz using a third-order Butterworth band-pass filter (as in [16, 18, 165]), and subsequently downsampled to $f_s = 64$ Hz. The envelopes of all considered 16 kHz speech signals were obtained using a Hilbert transform [129], followed by low-pass filtering at 8 Hz and downsampling to $f_s = 64$ Hz. For the training and decoding steps (see Section 2.3), the EEG recordings of each session were split into 10 trials, each of length 60 seconds (see Table 2.2). For filter training, the filter was computed using all considered trials based on (2.15) and (2.17), as proposed in [129,169], instead of computing a filter per trial and averaging per-trial filters as proposed in [16]. For filter training and evaluation, each participant's own data were used. The decoding performance was computed by averaging the percentage of correctly decoded trials over all considered trials and all participants.

2.5 Results and Discussion

In this section, the decoding performance of the least-squares-based AAD method is investigated for different acoustic conditions (see Table 2.2) using the experimental setup discussed in the previous section. Section 2.5.1 discusses the results of the questionnaire. In Section 2.5.2 the impact of different acoustic conditions for the training and decoding steps is investigated. In Section 2.5.3 the impact of the head filtering effect is explored by comparing the decoding performance using either the clean or the anechoic speech signals. Finally, in Section 2.5.4 the influence of each acoustic component is investigated by comparing the decoding performance using reference and training signals affected by background noise, reverberation, and/or interfering speaker.

2.5.1 *Questionnaire Analysis*

For all considered acoustic conditions, Fig. 2.2 presents the correct answer scores related to the attended story, averaged across all participants. The highest score is obtained for the anechoic condition, while the lowest score is obtained for the reverberant-noisy condition. The statistical multiple comparison test (Kruskal-Wallis test followed by the post-hoc Dunn and Sidak test [1]) showed a significant difference (Kruskal-Wallis test: $\chi^2 = 19.0$, $p = 0.002$) in terms of the correct answer score between the anechoic condition and either the noisy or the reverberant-noisy condition (post-hoc Dunn and Sidak test: $p = 0.022$ and $p = 0.000$, respectively) and between the reverberant condition and the reverberant-noisy condition (post-

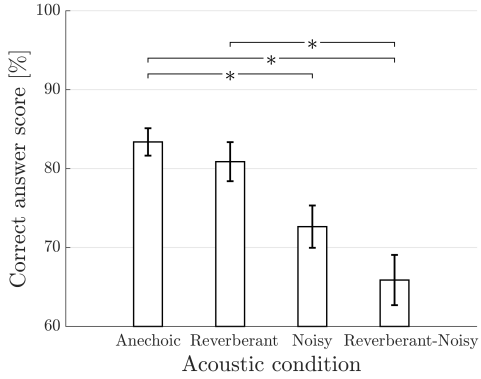


Fig. 2.2: The correct answer scores related to the attended story, averaged across all participants, for different acoustic conditions. Error bars represent one standard error around the mean and $*$ indicates a significant difference ($p < 0.05$) between acoustic conditions, based on the Kruskal-Wallis test followed by the post-hoc Dunn and Sidak test [1].

hoc Dunn and Sidak test: $p = 0.013$), implying that – as expected – the noisy and the reverberant-noisy condition are more challenging.

2.5.2 *Impact of Acoustic Conditions*

For all considered EEG evaluation conditions, Fig. 2.3 presents the decoding performance for different EEG training conditions when the clean speech signals are used as reference and training signals.

First, we investigate the feasibility of decoding EEG responses in different acoustic conditions $ec \in \{an, re, no, rn, ac\}$ when using filters trained using EEG responses in a *specific* acoustic condition $tc \in \{an, re, no, rn\}$ (i.e. left part of Fig. 2.3, separated by dashed line). When the EEG evaluation and training conditions are equal (indicated by ∇), it can be observed that a very good decoding performance ($> 96\%$) is obtained for all EEG evaluation conditions. These results are consistent with previous findings for the anechoic condition [19, 27, 164, 165, 180, 181] as well as with recent findings for the reverberant and the reverberant-noisy conditions [169]. For each EEG training condition $tc \in \{an, re, no, rn\}$, it can be observed that the decoding performance when the EEG evaluation and training conditions are equal (indicated by ∇) is among the highest decoding performances for all EEG evaluation conditions. When the EEG evaluation and training conditions are not equal, typically a lower decoding performance is obtained (except in some cases for the anechoic and the reverberant EEG training conditions). For example, for the reverberant-noisy EEG training condition the highest decoding performance is obtained for the reverberant-noisy EEG evaluation condition ($> 97\%$), while a lower decoding performance is obtained for the anechoic, reverberant, and noisy EEG evaluation conditions ($> 90\%$). In addition, for all EEG training conditions

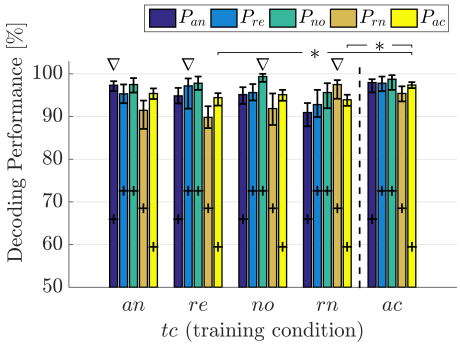


Fig. 2.3: The decoding performance for different EEG training and evaluation conditions when using the clean speech signals. The plus signs represent the upper boundary of the confidence interval corresponding to chance level, based on a binomial test at the 5% significance level, the error bars represent the bootstrap confidence interval at the 5% significance level, ∇ indicates the decoding performance when the EEG training and evaluation conditions are equal, and the dashed line separates the decoding performance obtained with filters trained in a specific acoustic condition and with filters trained in all acoustic conditions. A multiple comparison test (the Kruskal-Wallis test followed by the post-hoc Dunn and Sidak test) was performed across P_{ac} where $*$ indicates a significant difference ($p < 0.05$).

$tc \in \{an, re, no, rn\}$ it can be observed that the average decoding performance for all conditions P_{ac} is considerably high ($> 93\%$).

Secondly, we investigate the feasibility of decoding EEG responses in different acoustic conditions $ec \in \{an, re, no, rn, ac\}$ when using filters trained using EEG responses in *all acoustic conditions* $tc = ac$ (i.e. right part of Fig. 2.3, separated by dashed line). It can be observed that a very good decoding performance ($> 95\%$) is obtained for all EEG evaluation conditions and that the decoding performance across EEG evaluation conditions is more consistent compared to when using filters trained in a specific acoustic condition. In addition, the average decoding performance for all conditions P_{ac} obtained with filters trained in all conditions is occasionally significantly larger than with filters trained in a specific acoustic condition[1]. For example, the decoding performance P_{ac} obtained with filters trained in all conditions ($tc = ac$) is significantly larger than with filters trained either in the reverberant condition ($tc = re$) or in the reverberant-noisy condition ($tc = rn$) (Kruskal-Wallis test: $\chi^2 = 16.5$, $p = 0.002$; post-hoc Dunn and Sidak test comparisons of $tc = ac$ with $tc = re$ and $tc = rn$: $p = 0.020$, $p = 0.001$, respectively). To investigate how much the average decoding performance for all conditions P_{ac} varies across participants, Fig. 2.4 presents P_{ac} per participant, obtained with fil-

1 We also performed the AAD experiment using filters trained with 40 trials that were arbitrarily selected from different acoustic conditions and observed similar findings.

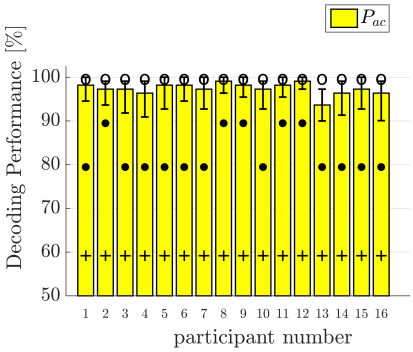


Fig. 2.4: The decoding performance P_{ac} per participant obtained with filters trained in all acoustic conditions and using clean speech signals. The plus signs represent the upper boundary of the confidence interval corresponding to chance level, based on a binomial test at the 5% significance level, the error bars represent the bootstrap confidence interval at the 5% significance level, the solid circles represent the minimum decoding performance and the void circles represent the maximum decoding performance.

ters trained in all conditions. It can be observed that the decoding performance per participant ranges between 80% and 100%.

The feasibility of using either filters trained in a specific acoustic condition or filters trained in all acoustic conditions to perform AAD in different acoustic conditions may be explained by considering the robust neural responses to degraded – but still intelligible – speech signals. Several studies have shown that auditory cortical responses resemble the clean attended speech signal more than the speech signal degraded by different acoustic components (e.g., background noise, interfering speaker), suggesting a robust neural representation of the clean attended speech signal [20, 152, 160, 175, 193]. To decode auditory attention, the trained filters aim at reconstructing the clean attended speech envelope from EEG responses that are largely invariant to degradations. Hence, the reconstructed attended envelope is expected to be more correlated to the clean attended speech envelope than to the clean unattended speech envelope, i.e. the correlation difference $(\rho^a - \rho^u)$ is expected to be larger than zero. For all considered EEG evaluation conditions, Fig. 2.5 presents the correlation difference for different EEG training conditions, averaged across all considered trials and participants (note that these average correlation coefficients are not directly used for decoding). It can be observed that a correlation difference significantly larger than zero is obtained for all considered acoustic conditions, which is consistent with a robust neural representation of the clean attended speech signal.

Finally, we investigate the parameters involved in the filter design (Δ, L, β) across EEG training conditions. Fig. 2.6 depicts the optimal parameter values (see Section 2.3.2), averaged across all considered trials and all participants. It can be observed that the optimal value for Δ varies only slightly between 93.8 ms to 101.6 ms,

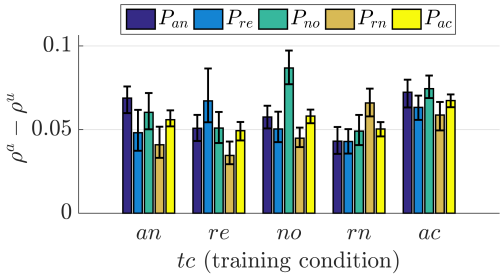


Fig. 2.5: Average correlation differences for different EEG training and evaluation conditions when using the clean speech signals. The error bars represent the bootstrap confidence interval at the 5% significance level.

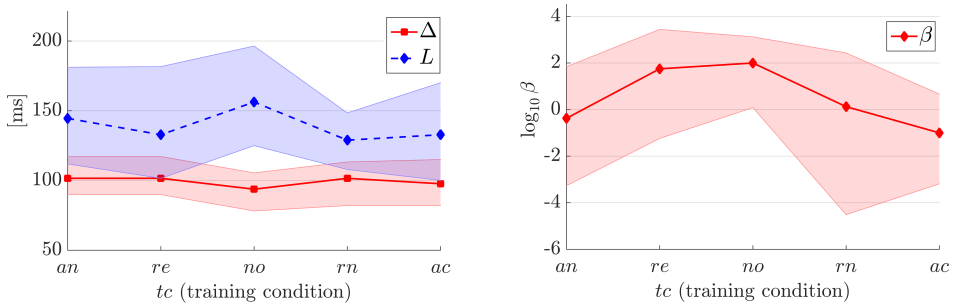


Fig. 2.6: The optimal values for the filter parameters (a) Δ and L, and (b) the regularization parameter β, averaged across all trials and all participants when using the clean speech signal. The shaded area indicates the bootstrap confidence interval at the 5% significance level.

while the optimal value for L varies more substantially between 109.3 ms to 128.9 ms. Accordingly, the EEG responses contributing most to the AAD performance are those with latencies between 93.8 ms and 230.5 ms, consistent with previous findings in [18, 19, 160]. In addition, the optimal value for the regularization parameter β varies between 10^{-1} to 10^{2}. It can be observed that the optimal regularization parameter is smaller when using filters trained in all conditions than when using filters trained in a specific acoustic condition. A possible explanation may be that training in all conditions can by itself be considered as some form of regularization, consistent with previous finding in [129].

In summary, the results in this section show the feasibility of using either filters trained in a specific acoustic condition or filters trained in all conditions to perform

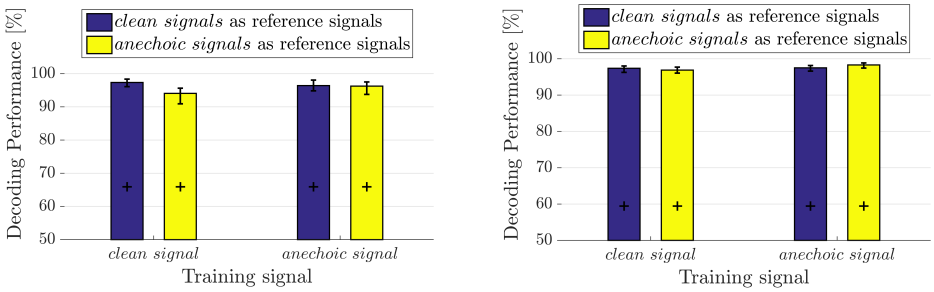


Fig. 2.7: Influence of head filtering effect on AAD. Comparison of decoding performance using either the clean or the anechoic speech signals when the EEG evaluation and training conditions are equal to (a) the anechoic condition or (b) all conditions. The plus signs represent the upper boundary of the confidence interval corresponding to chance level based on a binomial test at the 5% significance level, the error bars represent the bootstrap confidence interval at the 5% significance level.

AAD in different acoustic conditions[2]. While these results were obtained using the clean speech signals as training and reference signals, in the next sections we will investigate in more detail the influence of the different acoustic components (head filtering effect, reverberation, background noise, interfering speaker) in the training and reference signals.

2.5.3 *Influence of head filtering effect*

In this section, we investigate the influence of the head filtering effect by comparing the decoding performance when using clean or anechoic speech signals either as training or as reference signals. Fig. 2.7a presents the decoding performance for the anechoic condition ($ec = an$) when using filters trained in the anechoic condition ($tc = an$). Fig. 2.7b presents the average decoding performance for all conditions ($ec = ac$) when using filters trained in all conditions ($tc = ac$). A paired Wilcoxon signed rank test revealed no significant difference ($p > 0.05$) between using either the clean speech signals or the anechoic speech signals as training or as reference signals. These results indicate that for all considered acoustic conditions head filtering effects have no significant influence on the decoding performance.

2 We also performed the AAD experiment using trial lengths of 30 seconds and observed similar findings.

2.5.4 *Influence of background noise, reverberation and interfering speaker*

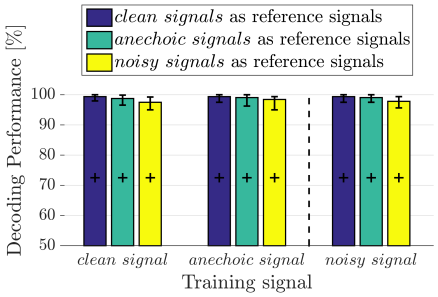
To investigate the influence of each acoustic component on AAD, Fig. 2.8 presents the decoding performance for all considered acoustic conditions (anechoic, reverberant, noisy, reverberant-noisy) using the following signals as training signals or as reference signals:

- the clean speech signals s^a and s^u.
- the anechoic speech signals $x_m^{a,an}$ and $x_m^{u,an}$.
- the anechoic speech signals affected by different acoustic components, i.e. the noisy speech signals $x_m^{a,no}$ and $x_m^{u,no}$ in (2.3) for the noisy condition, the reverberant speech signals x_m^a and x_m^u in (2.1) for the reverberant condition, the interfered speech signal x_m^{an} (attended and unattended side) in (2.2) for the anechoic condition[3], and the binaural speech signals y_m (attended and unattended side) in (2.1) for the reverberant-noisy condition.

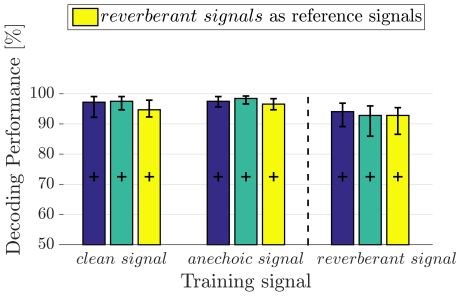
Similarly, Fig. 2.9 presents the correlation difference ($\rho^a - \rho^u$), averaged across all considered trials and participants (note that these average correlation coefficients are not directly used for decoding).

First, we investigate the case where the clean or the anechoic attended speech signal is used as training signal (i.e. left part of Fig. 2.8 and 2.9, separated by dashed line). When using the clean or anechoic speech signals as reference signals, a very good decoding performance ($> 94\%$) is obtained for all acoustic conditions, as already shown in Fig. 2.7. When using the noisy speech signals (in the noisy condition, Fig. 2.8a) or the reverberant speech signals (in the reverberant condition, Fig. 2.8b) as reference signals, there is no significant difference in decoding performance ($p > 0.05$) compared to when using the clean or anechoic speech signals as reference signals. On the other hand, when using the interfered speech signals (in the anechoic condition, Fig. 2.8c) or the binaural speech signals (in the reverberant-noisy condition, Fig. 2.8d) as reference signals, the decoding performance is significantly lower ($p < 0.05$) than when using the clean or anechoic speech signals as reference signals, although the decoding performance is still considerably large ($> 87\%$). The feasibility of using either the interfered speech signals or the binaural speech signals as reference signals for AAD can be explained by considering the broadband energy ratio between the attended and unattended speech components in the signals at the ears. As already mentioned in Section 2.2, due to the head filtering effect this broadband energy ratio is smaller at the side of the unattended speaker than at the side of the attended speaker. In summary, the results in Fig. 2.8 (left side) show that when using reference signals affected by reverberation or background noise, a comparable decoding performance can be obtained as when using clean or anechoic speech signals, whereas when using reference signals affected by the interfering speaker the decoding performance significantly decreases. This also suggests that in

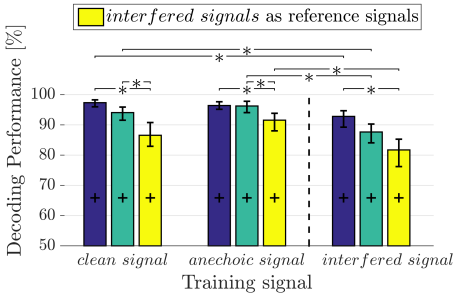
3 The interfered speech signal is used in the anechoic condition to exclude the influence of other acoustic components (background noise and reverberation) on the analysis.



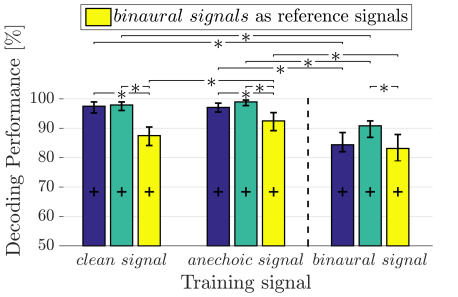
(a) Noisy condition ($ec = no$, $tc = no$)



(b) Reverberant condition ($ec = re$, $tc = re$)

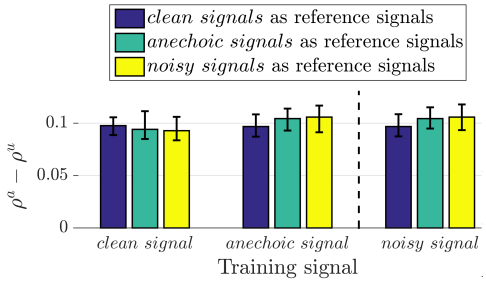


(c) Anechoic condition ($ec = an$, $tc = an$)

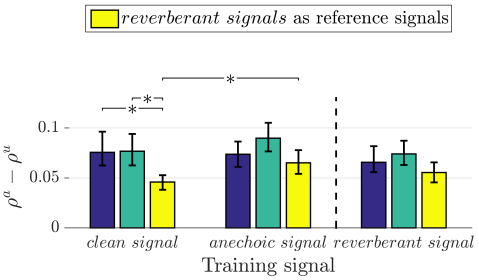


(d) Reverberant-noisy condition ($ec = rn$, $tc = rn$)

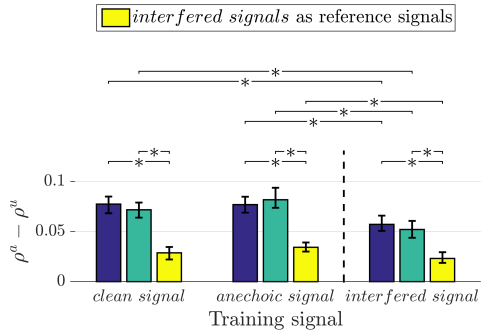
Fig. 2.8: Influence of different acoustic components (background noise, reverberation and interfering speaker) on AAD. Comparison of decoding performance when using (a) the noisy speech signals in the noisy condition, (b) the reverberant speech signals in the reverberant condition, (c) the interfered speech signals in the anechoic condition, (d) the binaural speech signals in the reverberant-noisy condition, either as training signal or as reference signals. The plus signs represent the upper boundary of the confidence interval corresponding to chance level based on a binomial test at the 5% significance level, and the error bars represent the bootstrap confidence interval at the 5% significance level. The dashed line separates the case where the clean or the anechoic attended speech signals are used as training signals and the case where the attended speech signals affected by different acoustic components are used as training signals. A paired Wilcoxon signed rank test was performed between the decoding performance using the clean or the anechoic speech signals as reference signals and using the anechoic speech signals affected by different acoustic components as reference signals. In addition, a paired Wilcoxon signed rank test was performed between the decoding performance using the clean or the anechoic speech signals as training signals and using the anechoic speech signals affected by different acoustic components as training signals. $*$ indicates a significant difference ($p < 0.05$) based on the paired Wilcoxon signed rank test.



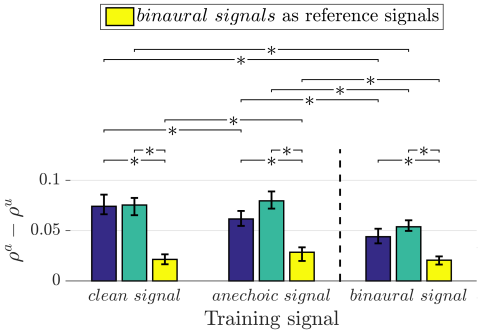
(a) Noisy condition ($ec = no$, $tc = no$)



(b) Reverberant condition ($ec = re$, $tc = re$)



(c) Anechoic condition ($ec = an$, $tc = an$)



(d) Reverberant-noisy condition ($ec = rn$, $tc = rn$)

Fig. 2.9: Influence of different acoustic components (background noise, reverberation and interfering speaker) on AAD. Comparison of correlation difference when using (a) the noisy speech signals in the noisy condition, (b) the reverberant speech signals in the reverberant condition, (c) the interfered speech signals in the anechoic condition, (d) the binaural speech signals in the reverberant-noisy condition, either as training signal or as reference signals. The error bars represent the bootstrap confidence interval at the 5% significance level, and $*$ indicates a significant difference ($p < 0.05$) based on the paired Wilcoxon signed rank test.

order to generate appropriate reference signals, it is more important to reduce the interfering speaker than to reduce background noise or reverberation.

The decoding performance results in Fig. 2.8 can be further explained by considering the influence of each acoustic component on the correlation difference in Fig. 2.9. For the noisy condition (Fig. 2.9a), there are no significant differences between the considered reference signals, which corresponds to the decoding performance results in Fig. 2.8a. For the reverberant condition (Fig. 2.9b), it can be observed that the correlation differences significantly decrease ($\rho^a - \rho^u < 0.04$) when using the reverberant speech signals as reference signals, but only when using the clean attended speech signal as training signal. Nevertheless, this lower correlation difference does not result in a significantly lower decoding performance in Fig. 2.8b. For the anechoic condition (Fig. 2.9c) and the reverberant-noisy condition (Fig. 2.9d), it can be observed that the correlation differences significantly decrease when using the interfered speech signals ($\rho^a - \rho^u < 0.03$) or the binaural speech signals ($\rho^a - \rho^u < 0.02$) as reference signals. These lower correlation differences are also reflected by significantly lower corresponding decoding performances in Fig. 2.8c and 2.8d.

Secondly, we explore the potential of using the attended speech signal affected by different acoustic components as training signal (i.e. right part of Fig. 2.8 and 2.9, separated by dashed line). On the one hand, when using the noisy attended speech signal (in the noisy condition, Fig. 2.8a) or the reverberant attended speech signal (in the reverberant condition, Fig. 2.8b) as training signal, there is no significant difference in decoding performance ($p > 0.05$) compared to when using the clean or the anechoic attended speech signal as training signal (for all considered reference signals). On the other hand, when using the interfered attended speech signal (in the anechoic condition, Fig. 2.8c) or the binaural attended speech signal (in the reverberant-noisy condition, Fig. 2.8d) as training signal, the decoding performance is significantly lower compared to when using either the clean or the anechoic attended speech signal as training signal (for all considered reference signals). Nevertheless, even when using the binaural attended speech signal as training signal in the reverberant-noisy condition, it is still feasible to perform AAD with a decoding performance larger than 82%. The decoding performance results in Fig. 2.8 when using attended speech signals affected by different acoustic components as training signal are mostly consistent with the correlation differences in Fig. 2.9.

In summary, the results in this section show that using speech signals affected by background noise and reverberation as training or reference signals results in a decoding performance that is comparable to using the clean or anechoic speech signals as training or reference signals. On the contrary, using speech signals affected by the interfering speaker as training or reference signals typically results in a significantly lower decoding performance. Table 2.3 presents which training/reference signals lead to the largest decoding performance for a specific EEG training/evaluation condition.

Table 2.3: Acoustic signal as training or reference signals using which the largest decoding performance for a specific EEG (training and evaluation) condition is obtained.

EEG Condition	Acoustic Signal
Noisy	Clean, anechoic and noisy signals
Reverberant	Clean, anechoic and reverberant signals
Anechoic	Clean and anechoic signals
Reverberant-noisy	Clean and anechoic signals

2.6 Conclusions

In this paper, we investigated the performance of the least-squares-based AAD method for different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy), both in the training step as well as in the decoding step. The experimental results showed that for all considered acoustic conditions it is possible to decode auditory attention with a considerably large decoding performance, even when the acoustic conditions for training and decoding are different. In addition, for most acoustic conditions there is no significant difference in decoding performance when using filters trained in all conditions or filters trained in a specific condition. This suggests that for an unseen realistic acoustic condition AAD can be performed using filters trained in, e.g., a laboratory acoustic condition.

Furthermore, we investigated the influence of the head filtering effect and of acoustic components (reverberation, background noise and interfering speaker) on the decoding performance. The experimental results showed that for all considered acoustic conditions the head filtering effect has no significant impact on the decoding performance. Moreover, when using speech signals affected by either reverberation or background noise as reference signals, a comparable decoding performance is obtained as when using clean speech signals as reference signals. On the contrary, when using speech signals affected by the interfering speaker as reference signals, the decoding performance significantly decreases. This suggests that for generating appropriate reference signals, e.g., using acoustic signal pre-processing algorithms, it is more important to reduce the interfering speaker than to reduce background noise or reverberation. Furthermore, when using the binaural speech signals as reference signals for decoding, a relatively large decoding performance can be obtained. This implies that decoding is feasible for the considered scenario even based on the unprocessed noisy and reverberant signals.

Finally, we explored the potential of using the attended speech signal affected by different acoustic components as training signal for computing the filter. When using attended speech signals affected by either reverberation or by background noise as training signal, a comparable decoding performance is obtained as when using the clean attended speech signal as training signal. However, when using attended speech signals affected by the interfering speaker as training signal, the decoding performance may significantly decrease. Nevertheless, even when using

the binaural attended speech signal as training signal, it is still feasible to achieve a large decoding performance.

While the discussion in this paper has been limited to the least-squares-based AAD method, in which auditory attention is decoded using an envelope reconstruction model, AAD approaches based on empirical mode decomposition [194,195], inherent fuzzy entropy [196] or a neural encoding model [24, 197] have not been investigated in this paper. Further work could therefore include a study on how reverberation and noise influence these AAD approaches.

3

COGNITIVE-DRIVEN BINAURAL BEAMFORMING USING EEG-BASED AUDITORY ATTENTION DECODING

Identifying the target speaker in hearing aid applications is an essential ingredient to improve speech intelligibility. Recently, a least-squares-based auditory attention decoding (AAD) method has been proposed to identify the target speaker from single-trial EEG recordings in an acoustic scenario with two competing speakers. Aiming at enhancing the target speaker and suppressing the interfering speaker and ambient noise, in this paper we propose a cognitive-driven speech enhancement system, consisting of a binaural beamformer which is steered based on AAD and estimated relative transfer function (RTF) vectors, which require estimates of the direction-of-arrivals (DOAs) of both speakers. For binaural beamforming and to generate reference signals for AAD, we consider either minimum-variance-distortionless-response (MVDR) beamformers or linearly-constrained-minimum-variance (LCMV) beamformers. Contrary to the binaural MVDR beamformer, the binaural LCMV beamformer allows to preserve the spatial impression of the acoustic scene and to control the suppression of the interfering speaker, which is important when intending to switch attention between speakers. The speech enhancement performance of the proposed system is evaluated in terms of the binaural signal-to-interference-plus-noise ratio (SINR) improvement in anechoic and reverberant conditions. Furthermore, we investigate the impact of RTF and DOA estimation errors and AAD errors on the speech enhancement performance. The experimental results show that the proposed system using LCMV beamformers yields a larger decoding performance and binaural SINR improvement compared to using MVDR beamformers.

3.1 Introduction

During the last decades significant advances have been made in multi-microphone speech enhancement algorithms for hearing aids. Although several algorithms are available to reduce background noise or to perform source separation in multi-talker scenarios [10, 13], their performance in improving speech intelligibility depends on correctly identifying the target speaker to be enhanced. In hearing aid applications, the target speaker is typically assumed to be either located in front of the listener

or to be the loudest speaker. However, since in real-world conditions these assumptions are often violated, the performance of speech enhancement algorithms may substantially decrease.

Recent advances have shown that it is possible to infer the auditory attention of a listener from electroencephalography (EEG) recordings [16,17]. Using single-trial EEG recordings, several auditory attention decoding (AAD) methods have been proposed to identify the attended speaker based on, e.g., a least-squares cost function [16], neural networks [23], and Bayesian filtering [24]. Aiming at incorporating AAD in a brain-computer interface for real-world applications, e.g., to control a hearing aid, a large research effort has recently focused on investigating the feasibility of AAD in real-world listening conditions [20,24–26,32,129,165,169,198], closing the loop of an AAD system by presenting feedback to the listener [180], and steering source separation and noise reduction algorithms based on AAD [27–29,166].

The least-squares-based AAD method proposed in [16] aims at reconstructing the attended speech envelope from the EEG recordings using a trained spatio-temporal filter. In the training step, the clean speech signal of the attended speaker is used to train the spatio-temporal filter by minimizing the least-squares error between the attended speech envelope and the reconstructed envelope. In the decoding step, the clean speech signals of both the attended and the unattended speaker are used as reference signals for decoding. Similarly to the least-squares-based AAD method, the AAD methods proposed in [23,24] use the clean speech signals of both speakers for decoding. In hearing aid applications, only the microphone signals, containing reverberation, background noise and interference, are obviously available in practice. In [26,169] it has been shown that AAD is still feasible using the noisy and reverberant microphone signals as reference signals, but the decoding performance is significantly decreased compared to using the clean speech signals as reference signals.

Aiming at generating appropriate reference signals for decoding from the microphone signals and incorporating AAD in speech enhancement algorithms, several cognitive-driven source separation and noise reduction algorithms [27–29,166,170] have been proposed. The single-microphone source separation algorithm proposed in [28] uses a deep neural network (DNN) to generate reference signals by separating the speakers from the mixture received at the microphone. Using electrocorticography recordings and AAD, one of the reference signals is then selected as the enhanced attended speaker. Although experimental results in [28] show that the algorithm is able to significantly improve the quality of the attended speaker, it should be realized that the algorithm is speaker-dependent, i.e., requires prior DNN training on known speakers. The multi-microphone noise reduction algorithms proposed in [27,29,166,170] are able to exploit the spatial diversity provided by the microphone signals for reference signal generation and speech enhancement. The cognitive-driven multi-channel Wiener filter (MWF) proposed in [27,166] generates reference signals from binaural hearing aid microphone signals using multiple multi-channel Wiener filters based on an envelope demixing algorithm and a voice activity detector (VAD). Using EEG recordings and AAD, one of the reference signals is then selected as the enhanced attended speaker. Experimental results in [27,166] show that the cognitive-driven MWF is able to enhance the attended speaker and

strongly suppress the interfering unattended speaker (especially in an anechoic condition). While strongly suppressing the interfering speaker is desired to improve speech intelligibility, it may deprive the listener from switching attention. In addition, the binaural MWF changes the spatial impression of the acoustic scene since all sources at the output of the binaural MWF are perceived as coming from the direction of the attended speaker [10, 95], which may lead to a confusion between acoustical and visual information.

Aiming at enhancing the attended speaker and controlling the suppression of the unattended speaker while preserving the spatial impression of the acoustic scene, in [170] a cognitive-driven speech enhancement system was proposed consisting of a binaural beamformer which is steered based on AAD and the estimated DOAs of the attended and the unattended speaker (see block diagram Fig. 3.1). First, the DOAs of both speakers are estimated from the binaural microphone signals. Based on the estimated DOAs, RTF vectors are selected from a database of (anechoic) prototype RTF vectors and two beamformers (BEAMs) generate reference signals for AAD. The least-squares-based AAD method then identifies the DOA of the attended and the unattended speaker to steer a binaural beamformer (BBEAM) for speech enhancement. To generate reference signals for AAD, in this paper we either consider minimum-variance-distortionless-response (MVDR) beamformers or a linearly-constrained-minimum-variance (LCMV) beamformers as BEAMs. While MVDR beamformers generate acceptable reference signals for decoding, we expect LCMV beamformers to generate better reference signals by jointly suppressing the interfering speaker and background noise. To generate binaural output signals, we either consider a steerable binaural MVDR beamformer or a steerable binaural LCMV beamformer as BBEAM. Contrary to the binaural MVDR beamformer, the binaural LCMV beamformer allows to control the suppression of the signal arriving from the unattended DOA and preserve the spatial impression of the acoustic scene. Compared to [170], in this paper we provide a detailed analysis and experimental comparison between the cognitive-driven binaural LCMV and MVDR beamformers for an acoustic scenario comprising two competing speakers and diffuse background noise in an anechoic and a reverberant condition. For the reverberant condition, we compare the performance between using (oracle or estimated) reverberant RTF vectors and anechoic prototype RTF vectors, which are determined by the (oracle or estimated) DOAs. In addition, we investigate the impact on the speech enhancement performance of RTF, DOA and AAD estimation errors and the STFT frame length. Moreover, we investigate how well the proposed cognitive-driven binaural beamformers preserve the spatial impression of the acoustic scene.

The paper is organized as follows. In Section 3.2 the configuration and the notation used for the binaural hearing aid setup recordings are introduced. In Section 3.3 the used DOA and RTF vector estimator are described. In Section 3.4 the beamformers and the AAD method used in the proposed cognitive-driven speech enhancement system are described. Section 3.5 describes the acoustic and EEG measurement setup, the algorithm implementation details and the performance measures. In Section 3.6 the experimental results are presented, exploring the decoding performance and the speech enhancement performance of the proposed system.

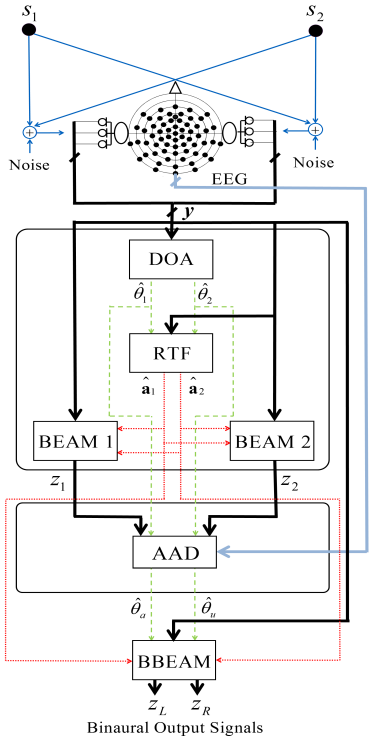


Fig. 3.1: Block diagram of the proposed cognitive-driven binaural beamformer in an acoustic scenario comprising two competing speakers (s_1 and s_2 with DOAs θ_1 and θ_2) and background noise. Based on the estimated DOAs ($\hat{\theta}_1$ and $\hat{\theta}_2$) of the speakers, the anechoic or reverberant RTF vectors ($\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$) are estimated and two beamformers (BEAM1 and BEAM2) generate reference signals (z_1 and z_2) for AAD. The AAD method then identifies the DOA of the attended and the unattended speaker ($\hat{\theta}_a$ and $\hat{\theta}_u$) to steer a binaural beamformer (BBEAM), generating binaural output signals z_L and z_R.

3.2 Configuration and notation

We consider an acoustic scenario comprising two competing speakers with DOAs θ_1 and θ_2 and background noise in a reverberant environment (see Fig. 3.1). The angle $\theta = 0^\circ$ corresponds to the frontal direction, while negative θ correspond to the left side of the listener and positive θ correspond to the right side. The clean signal of speaker 1 is denoted as $s_1[n]$, while the clean signal of speaker 2 is denoted as $s_2[n]$, with n the discrete time index. We consider a binaural hearing aid setup, where each hearing aid contains M microphones. The m-th microphone signal of the left hearing aid $y_{L,m}[n]$ can be decomposed as

$$y_{L,m}[n] = x_{1,L,m}[n] + x_{2,L,m}[n] + v_{L,m}[n], \tag{3.1}$$

where $x_{1,L,m}[n]$ and $x_{2,L,m}[n]$ denote the reverberant speech component in the m-th microphone signal corresponding to speaker 1 and speaker 2, respectively, and $v_{L,m}[n]$ denotes the background noise component. The reverberant speech components $x_{1,L,m}[n]$ and $x_{2,L,m}[n]$ consist of an anechoic speech component, encompassing the (anechoic) head filtering effect, and a reverberation component. The m-th microphone signal of the right hearing aid $y_{R,m}[n]$ can be decomposed similarly as in (3.1).

In the STFT domain, the $2M$-dimensional stacked vector of all microphone signals from the left and the right hearing aid is given by

$$\mathbf{y}(k,l) = [Y_{L,1}(k,l) \ \dots \ Y_{L,M}(k,l) \ \ Y_{R,1}(k,l) \ \dots \ Y_{R,M}(k,l)]^T, \tag{3.2}$$

where k denotes the frequency index and l denotes the frame index. For notational conciseness the indices k, l and n will be omitted in the remainder of this paper wherever possible.

Using (3.1) and (3.2), the signal vector $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{v}, \tag{3.3}$$

where the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{v}$ are defined similarly as in (3.2) for speaker 1, speaker 2, and the background noise, respectively. The vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are given by

$$\mathbf{x}_1 = \mathbf{h}_1 S_1, \ \mathbf{x}_2 = \mathbf{h}_2 S_2, \tag{3.4}$$

where $\mathbf{h}_1$ and $\mathbf{h}_2$ denote the $2M$-dimensional (reverberant) acoustic transfer function (ATF) vectors between the microphones on both hearing aids and speaker 1 and speaker 2, respectively. Using the first microphones on the left and the right hearing aid as so-called reference microphones, the vector $\mathbf{x}_1$ can be written as

$$\mathbf{x}_1 = \bar{\mathbf{a}}_{1,L} X_{1,L,1} = \bar{\mathbf{a}}_{1,R} X_{1,R,1}, \tag{3.5}$$

where $\bar{\mathbf{a}}_{1,\{L,R\}}$ denote the $2M$-dimensional relative transfer function (RTF) vectors [10, 13] of speaker 1 with respect to the reference microphones on the left and the right hearing aid, respectively. The RTF vectors $\bar{\mathbf{a}}_{2,\{L,R\}}$ of speaker 2 are defined similarly.

The output signals of all beamformers depicted in Fig. 3.1 are obtained by filtering and summing the microphone signals on both hearing aids, i.e.,

$$z = \mathrm{ISTFT}\left\{\mathbf{w}^H \mathbf{y}\right\}, \tag{3.6}$$

where $\mathrm{ISTFT}\{\cdot\}$ denotes the inverse short-time Fourier transform, $\mathbf{w}$ denotes the $2M$-dimensional filter vector, and $(\cdot)^H$ denotes the conjugate transpose operator.

3.3 DOA and RTF estimation

In this section, we present the algorithms to estimate the DOAs and the RTF vectors of both speakers, which will be used for beamforming and to generate reference

signals for AAD (see Section 3.4). Section 3.3.1 describes a classification-based DOA estimation algorithm. Section 3.3.2 describes two DOA-based RTF vector estimation algorithms.

3.3.1 *DOA estimation*

To estimate the DOAs of multiple speakers from binaural microphone signals, several methods have been proposed, e.g., by modeling binaural cues using a Gaussian mixture model [199], by using a beamforming-based approach [89], or by using a classification-based method [90]. In this paper, we will use the DOA estimation algorithm from [90], which estimates the source presence probability (SPP) for different DOAs using support vector machine (SVM) classifiers. The SVMs are trained to distinguish between the presence of a source for a certain direction and the absence for all other directions. The decision value of each SVM is mapped to the SPP $p_\theta[n]$ for each direction using a generalized linear model. As feature the short-term generalized cross-correlation with phase transform (GCC-PHAT) [200] is used, which has been shown to be relatively robust to noise and reverberation [201].

To estimate the DOAs of speakers 1 and 2, we first smooth the SPP for each direction across time, which increases the robustness against background noise, i.e.,

$$\overline{p}_\theta[n] = \tau p_\theta[n] + (1-\tau)\,\overline{p}_\theta[n-1], \tag{3.7}$$

with τ denoting the recursive smoothing constant. We then select two DOAs with the largest smoothed SPP $\overline{p}_\theta[n]$, from which the DOAs of speaker 1 and 2 are determined such that $\hat{\theta}_1 \leq \hat{\theta}_2$.

In the simulations (see Section 3.5), we will consider the following DOAs for speakers 1 and 2:

- ODOA: oracle DOAs, i.e., $\hat{\theta}_1 = \theta_1$ and $\hat{\theta}_2 = \theta_2$.
- EDOA: estimated DOAs $\hat{\theta}_1$ and $\hat{\theta}_2$ using [90] and (3.7).

3.3.2 *RTF vector estimation*

To estimate the RTF vectors $\bar{\mathbf{a}}_{1,\{L,R\}}$ and $\bar{\mathbf{a}}_{2,\{L,R\}}$ for both speakers, we will consider two approaches. In the first approach, the RTF vectors are approximated by *anechoic* RTF vectors $\mathbf{a}_L(\theta)$ and $\mathbf{a}_R(\theta)$, which are determined by the DOA θ (assuming the speakers are in the far field and in the horizontal plane). These anechoic RTF vectors can be either analytically computed based on a (spherical) head model, e.g., [109], or selected from a database of (measured) prototype RTF vectors, e.g., [110]. The estimated RTF vectors are denoted as $\mathbf{a}_{\{L,R\}}(\hat{\theta}_1)$ for speaker 1 and $\mathbf{a}_{\{L,R\}}(\hat{\theta}_2)$ for speaker 2.

The second approach aims at estimating the *reverberant* RTF vectors of both speakers directly from the microphone signals. Although many RTF vector estimation approaches are available for a single-speaker scenario [104,106,107,202], jointly estimating the RTF vectors of two simultaneously active speakers is not straightforward.

Assuming that the representation of both speakers in the STFT-domain is sparse, i.e., each time-frequency bin is dominated either by speaker 1 or speaker 2, we will first estimate the RTF vectors for each time-frequency bin assuming a single-speaker scenario and then assign these RTF vector estimates either to speaker 1 or speaker 2. The RTF vectors with respect to the reference microphones on the left and the right hearing aid are estimated as [203]

$$\tilde{\mathbf{a}}_{\{L,R\}}(k,l) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l)\,\mathbf{e}_{\{L,R\}}}{\mathbf{e}_{\{L,R\}}^{T}\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l)\,\mathbf{e}_{\{L,R\}}}, \tag{3.8}$$

with the smoothed microphone covariance matrix at each time-frequency bin computed as

$$\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l) = \alpha \boldsymbol{y}(k,l)\,\boldsymbol{y}^{H}(k,l) + (1-\alpha)\,\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l-1), \tag{3.9}$$

with α recursive smoothing constant and $\mathbf{e}_{\{L,R\}}$ reference microphone selection vectors consisting of zeros and one element equal to 1, i.e., $\mathbf{e}_L(1) = 1$ and $\mathbf{e}_R(M+1) = 1$. The estimated RTF vectors $\tilde{\mathbf{a}}_{\{L,R\}}(k,l)$ are then assigned to either speaker 1 or 2 based on their corresponding DOA and the estimated DOAs of both speakers. Similarly as in [204], the corresponding DOA $\hat{\theta}_{\mathrm{RTF}}(k,l)$ of the estimated RTF vectors is determined per time-frequency bin by computing the normalized cross-correlation $\hat{\kappa}(k,l)$ between the reference microphone signals (corresponding to the phase difference) with the normalized cross-correlation $\kappa(\theta)$ between the anechoic prototype RTF vectors for all directions θ, i.e.,

$$\hat{\theta}_{\mathrm{RTF}}(k,l) = \underset{\theta}{\operatorname{argmin}}(|\hat{\kappa}(k,l) - \kappa(\theta)|), \tag{3.10}$$

with

$$\hat{\kappa}(k,l) = \frac{\mathbf{e}_L^{T}\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l)\,\mathbf{e}_R}{\left|\mathbf{e}_L^{T}\hat{\mathbf{\Phi}}_{\mathbf{y}}(k,l)\,\mathbf{e}_R\right|}, \tag{3.11}$$

$$\kappa(\theta) = \frac{\mathbf{e}_R^{T}\mathbf{a}_L(\theta)}{\left|\mathbf{e}_R^{T}\mathbf{a}_L(\theta)\right|}. \tag{3.12}$$

When the DOA $\hat{\theta}_{\mathrm{RTF}}(k,l)$ is in a region of $\pm 5^{\circ}$ around the estimated DOA $\hat{\theta}_1(l)$ of speaker 1, then the estimated RTF vectors $\tilde{\mathbf{a}}_{\{L,R\}}(k,l)$ are assigned to speaker 1 and recursively smoothed, i.e.,

$$\hat{\mathbf{a}}_{1,\{L,R\}}(k,l) = \beta\tilde{\mathbf{a}}_{\{L,R\}}(k,l) + (1-\beta)\,\hat{\mathbf{a}}_{1,\{L,R\}}(k,l-1). \tag{3.13}$$

with β recursive smoothing constant. When the DOA $\hat{\theta}_{\mathrm{RTF}}(k,l)$ is in a region of $\pm 5^{\circ}$ around the estimated DOA $\hat{\theta}_2(l)$ of speaker 2, then the estimated RTF vectors $\tilde{\mathbf{a}}_{\{L,R\}}(k,l)$ are assigned to speaker 2 and recursively smoothed, i.e.,

$$\hat{\mathbf{a}}_{2,\{L,R\}}(k,l) = \beta\tilde{\mathbf{a}}_{\{L,R\}}(k,l) + (1-\beta)\,\hat{\mathbf{a}}_{2,\{L,R\}}(k,l-1). \tag{3.14}$$

3.4 Cognitive-driven binaural beamformer

In this section, we present the proposed cognitive-driven speech enhancement system (see Fig. 3.1), consisting of three main blocks. Section 3.4.1 describes the reference signal generation, where either MVDR or LCMV beamformers generate reference signals for AAD using the estimated DOA-based RTF vectors. Section 3.4.2 reviews the least-squares-based AAD method in [16], which is used to identify the DOA of the attended and the unattended speaker. These DOAs are used to steer a binaural MVDR or LCMV beamformer generating binaural output signals, which is discussed in Section 3.4.3.

3.4.1 *Reference signal generation using beamformers*

In [26, 32] it has been shown that the decoding performance of the least-squares-based AAD method (see Section 3.4.2) is heavily affected by the presence of background noise and especially the interfering speaker in the reference signals used for decoding. In this paper we will investigate different beamformers for generating appropriate reference signals from the binaural microphone signals.

In [29] it has been proposed to use an MVDR beamformer for generating reference signals. The MVDR beamformer [9,10,86] using RTF vectors $\bar{\mathbf{a}}$ aims at minimizing the power spectral density (PSD) of the output noise component while preserving the target speech component in one of the microphone signals. The corresponding constrained optimization problem is given by

$$\min_{\mathbf{w}} \underbrace{\mathbf{w}^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{w}}_{\text{noise output PSD}} \quad \text{subject to} \quad \underbrace{\mathbf{w}^H \bar{\mathbf{a}}_t = 1}_{\text{target}}, \tag{3.15}$$

where $\mathbf{\Phi}_{\mathbf{v}} = \varepsilon\left\{\mathbf{v}\mathbf{v}^H\right\}$ denotes the noise covariance matrix with $\varepsilon\{\cdot\}$ the expected value operator, and $\bar{\mathbf{a}}_t$ denotes the RTF vector corresponding to the target speaker. The MVDR beamformer solving (3.15) is given by [9,10,86]

$$\mathbf{w}_{\text{MVDR}} = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1} \bar{\mathbf{a}}_t}{\bar{\mathbf{a}}_t^H \mathbf{\Phi}_{\mathbf{v}}^{-1} \bar{\mathbf{a}}_t}. \tag{3.16}$$

A disadvantage of the MVDR beamformer is the fact that an interfering speaker may not be sufficiently suppressed, possibly reducing the AAD performance. Hence, to jointly suppress the interfering speaker and background noise, we will also consider the LCMV beamformer [9, 205], which adds an interference suppression constraint to the MVDR optimization problem in (3.15), i.e.,

$$\min_{\mathbf{w}} \underbrace{\mathbf{w}^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{w}}_{\text{noise output PSD}} \quad \text{subject to} \quad \underbrace{\mathbf{w}^H \bar{\mathbf{a}}_t = 1}_{\text{target}}, \ \underbrace{\mathbf{w}^H \bar{\mathbf{a}}_i = 0}_{\text{interference}}, \tag{3.17}$$

where $\bar{\mathbf{a}}_i$ denotes the RTF vector corresponding to the interfering speaker. The LCMV beamformer solving (3.17) is given by [205, 206]

$$\mathbf{w}_{\text{LCMV}} = \mathbf{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}\left(\bar{\mathbf{C}}^H\mathbf{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}\right)^{-1}\mathbf{b}, \tag{3.18}$$

with

$$\bar{\mathbf{C}} = \left[\bar{\mathbf{a}}_t \quad \bar{\mathbf{a}}_i\right], \quad \mathbf{b} = \left[1 \quad 0\right]^T. \tag{3.19}$$

Since in practice it is not trivial to accurately estimate both RTF vectors $\bar{\mathbf{a}}_t$ and $\bar{\mathbf{a}}_i$ in a noisy and reverberant environment, in this paper we will also consider beamformers using anechoic RTF vectors $\mathbf{a}(\theta)$. The MVDR beamformer in (3.16) with target angle θ_t is given by

$$\mathbf{w}_{\text{MVDR}}(\theta_t) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}(\theta_t)}{\mathbf{a}^H(\theta_t)\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}(\theta_t)}. \tag{3.20}$$

The LCMV beamformer in (3.18) with target angle θ_t and interfering angle θ_i is given by

$$\mathbf{w}_{\text{LCMV}}(\theta_t, \theta_i) = \mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}\right)^{-1}\mathbf{b}, \tag{3.21}$$

with

$$\mathbf{C} = \left[\mathbf{a}(\theta_t) \quad \mathbf{a}(\theta_i)\right], \quad \mathbf{b} = \left[1 \quad 0\right]^T. \tag{3.22}$$

Aiming at generating appropriate reference signals, i.e., separated speaker signals with reduced noise, we will either use two MVDR beamformers or two LCMV beamformers (BEAM1 and BEAM2 in Fig. 3.1), employing either anechoic or reverberant RTF vectors, i.e.,

- MVDR beamformers: an MVDR beamformer with estimated target angle $\theta_t = \hat{\theta}_1$ to generate the reference signal for speaker 1, and an MVDR beamformer with estimated target angle $\theta_t = \hat{\theta}_2$ to generate the reference signal for speaker 2.
- LCMV beamformers: an LCMV beamformer with estimated target angle $\theta_t = \hat{\theta}_1$ and estimated interfering angle $\theta_i = \hat{\theta}_2$ to generate the reference signal for speaker 1, and an LCMV beamformer with estimated target angle $\theta_t = \hat{\theta}_2$ and estimated interfering angle $\theta_i = \hat{\theta}_1$ to generate the reference signal for speaker 2.

It should be noted that we consider the RTF vectors normalized with respect to the left microphone (i.e., $\mathbf{a}_L(\hat{\theta}_{\{t,i\}})$ and $\hat{\mathbf{a}}_{\{t,i\},L}$) if $\hat{\theta}_t \leq 0°$, and normalized with respect to the right microphone (i.e., $\mathbf{a}_R(\hat{\theta}_{\{t,i\}})$ and $\hat{\mathbf{a}}_{\{t,i\},R}$) if $\hat{\theta}_t > 0°$.
The output signals of the MVDR and LCMV beamformers can be decomposed as

$$z_1 = z_{t,1} + z_{i,1} + z_{v,1}, \tag{3.23}$$

$$z_2 = z_{t,2} + z_{i,2} + z_{v,2}, \tag{3.24}$$

where $z_{t,1}$ and $z_{t,2}$ denote the target output speech component, $z_{i,1}$ and $z_{i,2}$ denote the interfering output speech component, and $z_{v,1}$ and $z_{v,2}$ denote the output noise component.

In the simulations (see Section 3.5), we will consider the following RTF vectors for the MVDR and LCMV beamformers:

- ORTF: oracle reverberant RTF vectors $\bar{\mathbf{a}}_1$ and $\bar{\mathbf{a}}_2$.
- ODOA: anechoic RTF vectors $\mathbf{a}(\theta)$ using the oracle DOAs θ_1 and θ_2.
- ERTF: estimated RTF vectors $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$
- EDOA: anechoic RTF vectors $\mathbf{a}(\theta)$ using the estimated DOAs $\hat{\theta}_1$ and $\hat{\theta}_2$.

It should be noted that for the anechoic condition the oracle RTF vectors (ORTF) are obviously the same as the anechoic RTF vectors using the oracle DOAs (ODOA).

3.4.2 *Auditory attention decoding*

Based on the reference signals z_1 and z_2 generated by the MVDR or LCMV beamformers, the EEG-based auditory attention decoding method then aims at identifying which speaker the listener attended to. This section briefly describes the least-squares-based AAD method from [16], which consists of a training and a decoding step.

3.4.2.1 *Decoding step*

The EEG recordings are first segmented into trials (see Section 3.5.2 for more details). To decode auditory attention from C-channel EEG recordings $r_c[i]$, with $c = 1 \dots C$ and i the sub-sampled time index of a trial, it has been proposed in [16] to reconstruct an estimate of the attended speech envelope $\hat{e}_a[i]$ using a trained spatio-temporal filter, i.e.,

$$\hat{e}_a[i] = \mathbf{g}^T \mathbf{r}[i], \; i = 1 \dots I, \tag{3.25}$$

with

$$\mathbf{g} = \left[\mathbf{g}_1^T \, \mathbf{g}_2^T \dots \, \mathbf{g}_C^T\right]^T, \tag{3.26}$$

$$\mathbf{g}_c = \left[g_{c,0} \; g_{c,1} \dots \, g_{c,J-1}\right]^T, \tag{3.27}$$

$$\mathbf{r}[i] = \left[\mathbf{r}_1^T[i] \, \mathbf{r}_2^T[i] \; \dots \; \mathbf{r}_C^T[i]\right]^T, \tag{3.28}$$

$$\mathbf{r}_c[i] = \left[r_c[i+\Delta] \; r_c[i+1+\Delta] \; \dots \; r_c[i+J-1+\Delta]\right]^T, \tag{3.29}$$

where J denotes the number of filter coefficients per channel and Δ models the latency of the attentional effect in the EEG responses to acoustic stimuli. Next, the correlation coefficients between the estimated attended speech envelope $\hat{e}_a[i]$ in (3.25) and the envelope of two reference signals are computed as

$$\rho_1 = \rho\left(e_1[i], \hat{e}_a[i]\right), \; \rho_2 = \rho\left(e_2[i], \hat{e}_a[i]\right), \tag{3.30}$$

where $e_1[i]$ and $e_2[i]$ denote the envelopes of the reference signals z_1 and z_2, respectively, such that ρ_1 and ρ_2 denote the correlation coefficients corresponding to speaker 1 and speaker 2, respectively. Based on these correlation coefficients, it is then decided that the listener attended to speaker 1 if $\rho_1 > \rho_2$ or attended to speaker 2 otherwise.

In this paper, we hence propose to estimate the DOA of the attended speaker θ_a and the DOA of unattended speaker θ_u based on the correlation coefficients ρ_1 and ρ_2 and the estimated DOAs of speaker 1 and 2 as

$$\begin{cases} \hat{\theta}_a = \hat{\theta}_1, \quad \hat{\theta}_u = \hat{\theta}_2 & \text{if} \quad \rho_1 > \rho_2 \\ \hat{\theta}_a = \hat{\theta}_2, \quad \hat{\theta}_u = \hat{\theta}_1 & \text{otherwise.} \end{cases} \tag{3.31}$$

To investigate the impact of AAD errors on the speech enhancement performance of the proposed system, in the simulations we will consider either

- oracle AAD (OAAD), i.e., $\hat{\theta}_a = \theta_a$ and $\hat{\theta}_u = \theta_u$
- estimated AAD (EAAD), where $\hat{\theta}_a$ and $\hat{\theta}_u$ are determined using (3.31).

3.4.2.2 *Training step*

Prior to the decoding step, the spatio-temporal filter $\mathbf{g}$ in (3.25) needs to be trained. During the training step the attended speaker is obviously assumed to be known. The filter $\mathbf{g}$ is computed by minimizing the least-squares error between the attended speech envelope $e_a[i]$ and the reconstructed envelope $\hat{e}_a[i]$, regularized with the squared l_2-norm of the derivatives of the filter coefficients to avoid over-fitting [16, 26, 129, 169], i.e.,

$$\min_{\mathbf{g}} \frac{1}{I} \sum_{i=1}^{I} \left(e_a[i] - \mathbf{g}^T \mathbf{r}[i] \right)^2 + \beta \mathbf{g}^T \mathbf{D} \mathbf{g}, \tag{3.32}$$

with $\mathbf{D}$ denoting the derivative matrix [169] and β denoting a regularization parameter. The filter minimizing the regularized least-squares cost function in (3.32) is equal to

$$\mathbf{g} = \left(\mathbf{Q} + \beta \mathbf{D} \right)^{-1} \mathbf{q}, \tag{3.33}$$

with the correlation matrix $\mathbf{Q}$ and the cross-correlation vector $\mathbf{q}$ given by

$$\mathbf{Q} = \frac{1}{I} \sum_{i=1}^{I} \left(\mathbf{r}[i] \, \mathbf{r}^T[i] \right), \quad \mathbf{q} = \frac{1}{I} \sum_{i=1}^{I} \left(\mathbf{r}[i] \, e_a[i] \right). \tag{3.34}$$

3.4.3 *Binaural beamformer*

The estimated DOAs of the attended and the unattended speaker in (3.31) are then used to steer a binaural beamformer (BBEAM in Fig. 3.1), generating binaural output signals. For binaural beamforming, we will consider either the binaural MVDR or LCMV beamformer.

The binaural MVDR beamformer [95] aims at minimizing the PSD of the output noise component while passing signals arriving from the estimated attended DOA $\hat{\theta}_a$. Similarly to (3.16), the binaural MVDR beamformer for the left and the right hearing aid using the estimated RTF vectors is given by

$$\mathbf{w}_{\text{BMVDR,L}} = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\hat{\mathbf{a}}_{a,L}}{\hat{\mathbf{a}}_{a,L}^{H}\mathbf{\Phi}_{\mathbf{v}}^{-1}\hat{\mathbf{a}}_{a,L}}, \tag{3.35}$$

$$\mathbf{w}_{\text{BMVDR,R}} = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\hat{\mathbf{a}}_{a,R}}{\hat{\mathbf{a}}_{a,R}^{H}\mathbf{\Phi}_{\mathbf{v}}^{-1}\hat{\mathbf{a}}_{a,R}}, \tag{3.36}$$

where $\hat{\mathbf{a}}_{a,\{L,R\}} = \hat{\mathbf{a}}_{1,\{L,R\}}$ if $\hat{\theta}_a = \hat{\theta}_1$ or $\hat{\mathbf{a}}_{a,\{L,R\}} = \hat{\mathbf{a}}_{2,\{L,R\}}$ if $\hat{\theta}_a = \hat{\theta}_2$. When using anechoic RTF vectors, the binaural MVDR beamformer for the left and the right hearing aid is given by

$$\mathbf{w}_{\text{BMVDR,L}}(\hat{\theta}_a) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}_L(\hat{\theta}_a)}{\mathbf{a}_L^{H}(\hat{\theta}_a)\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}_L(\hat{\theta}_a)}, \tag{3.37}$$

$$\mathbf{w}_{\text{BMVDR,R}}(\hat{\theta}_a) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}_R(\hat{\theta}_a)}{\mathbf{a}_R^{H}(\hat{\theta}_a)\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}_R(\hat{\theta}_a)}. \tag{3.38}$$

It should be noted that the binaural MVDR beamformer preserves the binaural cues, i.e., the interaural level difference (ILD) and the interaural time difference (ITD), of the signals arriving from the attended DOA, but distorts the binaural cues of signals arriving from other directions (including background noise). Since all sources are perceived as coming from the direction of the attended speaker, this will change the spatial impression of the acoustic scene.

As an alternative to the binaural MVDR beamformer, we also consider the binaural LCMV beamformer [91, 122], which allows to control the suppression of signals arriving from the unattended DOA (possibly enabling the listener to switch attention) and preserves the binaural cues of signals arriving from the attended and the unattended DOA. The binaural LCMV beamformer aims at minimizing the PSD of the output noise component while passing signals arriving from the estimated attended DOA $\hat{\theta}_a$ and suppressing signals arriving from the estimated unattended DOA $\hat{\theta}_u$. Similarly to (3.18), the binaural LCMV beamformer for the left and the right hearing aid using the estimated RTF vectors is given by

$$\mathbf{w}_{\text{BLCMV,L}} = \mathbf{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}_L\left(\bar{\mathbf{C}}_L^{H}\mathbf{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}_L\right)^{-1}\mathbf{b}_L, \tag{3.39}$$

$$\mathbf{w}_{\text{BLCMV,R}} = \mathbf{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}_R\left(\bar{\mathbf{C}}_R^{H}\mathbf{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}_R\right)^{-1}\mathbf{b}_R, \tag{3.40}$$

with

$$\bar{\mathbf{C}}_L = [\hat{\mathbf{a}}_{a,L} \quad \hat{\mathbf{a}}_{u,L}], \quad \mathbf{b}_L = [1 \quad \delta_L]^T, \tag{3.41}$$

$$\bar{\mathbf{C}}_R = [\hat{\mathbf{a}}_{a,R} \quad \hat{\mathbf{a}}_{u,R}], \quad \mathbf{b}_R = [1 \quad \delta_R]^T, \tag{3.42}$$

where $\hat{\mathbf{a}}_{u,\{L,R\}} = \hat{\mathbf{a}}_{1,\{L,R\}}$ if $\hat{\theta}_u = \hat{\theta}_1$ or $\hat{\mathbf{a}}_{u,\{L,R\}} = \hat{\mathbf{a}}_{2,\{L,R\}}$ if $\hat{\theta}_u = \hat{\theta}_2$, and $0 \leq \delta_L \leq 1$ and $0 \leq \delta_R \leq 1$ denote the interference suppression factors for the left and the right hearing aid, respectively. When using anechoic RTF vectors, the binaural LCMV beamformer for the left and the right hearing aid is given by

$$\mathbf{w}_{\text{BLCMV,L}}(\hat{\theta}_a, \hat{\theta}_u) = \mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}_L \left(\mathbf{C}_L^H \mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}_L\right)^{-1} \mathbf{b}_L, \tag{3.43}$$

$$\mathbf{w}_{\text{BLCMV,R}}(\hat{\theta}_a, \hat{\theta}_u) = \mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}_R \left(\mathbf{C}_R^H \mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}_R\right)^{-1} \mathbf{b}_R, \tag{3.44}$$

with

$$\mathbf{C}_L = \left[\mathbf{a}_L(\hat{\theta}_a) \quad \mathbf{a}_L(\hat{\theta}_u)\right], \tag{3.45}$$

$$\mathbf{C}_R = \left[\mathbf{a}_R(\hat{\theta}_a) \quad \mathbf{a}_R(\hat{\theta}_u)\right]. \tag{3.46}$$

Since we aim at preserving the spatial impression of the acoustic scene, we will use the same interference suppression factor for the left and the right hearing aid, i.e., $\delta = \delta_L = \delta_R$. Setting δ to zero corresponds to a complete suppression of signals arriving from the estimated unattended DOA $\hat{\theta}_u$ but unpredictable binaural cue distortion for the unattended speaker (due to using anechoic RTF vectors in a reverberant environment or DOA estimation errors in an anechoic environment), while $\delta > 0$ leads to a more controlled suppression and binaural cue preservation of the unattended speaker [91,122]. The output signals of the binaural LCMV beamformer for the left and the right hearing aid are equal to

$$z_L = \text{ISTFT}\left\{\mathbf{w}_{\text{BLCMV,L}}^H(\hat{\theta}_a, \hat{\theta}_u)\boldsymbol{y}\right\}, \tag{3.47}$$

$$z_R = \text{ISTFT}\left\{\mathbf{w}_{\text{BLCMV,R}}^H(\hat{\theta}_a, \hat{\theta}_u)\boldsymbol{y}\right\}. \tag{3.48}$$

The binaural output signals of the binaural MVDR and LCMV beamformers can be decomposed as

$$z_L = z_{a,L} + z_{u,L} + z_{v,L}, \tag{3.49}$$

$$z_R = z_{a,R} + z_{u,R} + z_{v,R}, \tag{3.50}$$

where $z_{a,L}$ and $z_{a,R}$ denote the (oracle) attended output speech component, $z_{u,L}$ and $z_{u,R}$ denote the (oracle) unattended output speech component, and $z_{v,L}$ and $z_{v,R}$ denote the output noise component.

3.5 Experimental setup

In this section, we describe the acoustic simulation setup, the setup used for EEG measurements, AAD training and evaluation, the implementation details for the DOA estimation, RTF estimation and beamforming algorithms, and the used performance measures.

Table 3.1: Reference signals used for evaluating the AAD performance

Reference Signals	Abbreviation	Description
Oracle anechoic signals	ORACLE	Anechoic speech components of the attended and the unattended speaker
Processed microphone signals	ORTF	Output signals of MVDR/LCMV beamformers using oracle reverberant RTFs
	ODOA	Output signals of MVDR/LCMV beamformers using anechoic RTFs and oracle DOAs
	ERTF	Output signals of MVDR/LCMV beamformers using estimated reverberant RTFs
	EDOA	Output signals of MVDR/LCMV beamformers using anechoic RTFs and estimated DOAs
Unprocessed microphone signals	UNPROC	Noisy and reverberant reference microphone signals

3.5.1 *Acoustic simulation setup*

Two German audio stories, uttered by two different male speakers, were used as the clean speech signals s_1 and s_2. Speech pauses from the audio stories that exceeded 0.5 s were shortened to 0.5 s, resulting in two highly overlapping (competing) audio stories. The hearing aid microphone signals $y_{L,m}$ and $y_{R,m}$ were generated at a sampling frequency of 16 kHz by convolving the clean speech signals with non-individualized measured binaural impulse responses (anechoic or reverberant) for a binaural hearing aid setup from [110], and adding diffuse babble noise. The diffuse babble noise was simulated according to [192] using babble speech recordings and a cylindrically isotropic noise field assumption. The hearing aid setup in [110] consisted of two hearing aids, each equipped with 3 microphones, mounted on a dummy head. As reference microphones, we chose the front microphones of the left and the right hearing aid. The left and the right competing speaker were simulated at $\theta_1 = -45^\circ$ and $\theta_2 = 45^\circ$ (see Fig. 3.2). In total, four acoustic conditions were considered: two anechoic conditions with binaural input SNRs 9.0 dB and 4.0 dB, and two reverberant conditions (reverberation time $T_{60} \approx 0.5$ s) with the same SNRs. The binaural input SNR is defined as the energy ratio between the speech components of speaker 1 and 2 in the reference microphone signals and the background noise components in the reference microphone signals, i.e.,

$$\mathrm{BSNR}_{in} = 10\log_{10}\frac{\phi_x}{\phi_v}, \tag{3.51}$$

with

$$\begin{aligned}\phi_x &= \varepsilon\left\{|x_{1,L,1}|^2\right\} + \varepsilon\left\{|x_{2,L,1}|^2\right\} + \\ &\quad \varepsilon\left\{|x_{1,R,1}|^2\right\} + \varepsilon\left\{|x_{2,R,1}|^2\right\}, \\ \phi_v &= \varepsilon\left\{|v_{L,1}|^2\right\} + \varepsilon\left\{|v_{R,1}|^2\right\}.\end{aligned} \tag{3.52}$$

3.5.2 *EEG measurement and AAD setup*

Eighteen normal-hearing and German-speaking participants took part in this study (see [26]). As acoustic stimuli, the reference microphone signals of the left and the

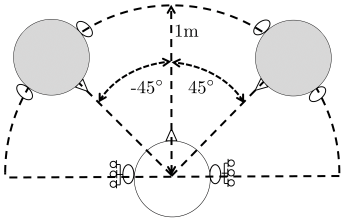


Fig. 3.2: Acoustic simulation setup for the reverberant condition. Two competing speakers were located at DOAs $\theta_1 = -45°$ and $\theta_2 = 45°$ and a distance of 1 m from the listener with two hearing aids, each equipped with 3 microphones.

right hearing aid were presented to the participants via insert earphones (E-A-RTONE 3A)[1]. Among all participants, 8 participants were instructed to attend to the left speaker, while 10 participants were instructed to attend to the right speaker. Two participants were excluded from the analysis, one participant due to poor attentional performance and the other one due to a technical hardware issue.

For all acoustic conditions, the EEG responses $r_c[i]$ were recorded using $C = 64$ channels[2] at a sampling frequency of 500 Hz, and referenced to the nose electrode (see [26] for more details). Similarly as in [16, 26], the EEG responses were re-referenced offline to a common average reference, band-pass filtered between 2 Hz and 8 Hz using a third-order Butterworth band-pass filter, and subsequently down-sampled to 64 Hz.

For filter training and decoding (see Section 3.4.2), the attended speech envelope $e_a[i]$ as well as the envelopes $e_1[i]$ and $e_2[i]$ of the reference signals were obtained using a Hilbert transform [129], followed by low-pass filtering at 8 Hz and down-sampling to 64 Hz. The attended speech envelope $e_a[i]$ was computed from the anechoic speech component of the attended speaker in the reference microphone signal at the side of the attended speaker. The EEG recordings for the different acoustic conditions were grouped together based on reverberation time, resulting in two experimental analysis conditions, i.e., anechoic and reverberant. The EEG recordings corresponding to each experimental analysis condition were split into 40 trials, each of length 30 seconds. Each participant's own data were used for filter training and evaluation. To avoid using the same trial for filter training and decoding, the leave-one-out cross validation approach was used (see [26] for more details). All analyses were performed using the EEG recordings under the same experimental analysis condition. Similarly as in [29], the parameters of the spatio-temporal filter $\mathbf{g}$ in (3.25) were set to $J = 8$ and $\Delta = 8$ (corresponding to 125 ms).

The AAD performance will be evaluated for both experimental analysis conditions using several reference signals z_1 and z_2 (see Table 3.1):

1 Please note that during the EEG measurement the participants were only presented the reference microphone signals, not the binaural output signals of the proposed system.

2 BrainCap with multitrodes from Easycap GmbH.

- oracle anechoic signals, corresponding to perfectly separated speech signals of the attended and the unattended speaker, i.e., the anechoic speech component of the attended speaker in the reference microphone signal at the side of the attended speaker and the anechoic speech component of the unattended speaker in the reference microphone signal at the side of the unattended speaker.
- processed microphone signals, i.e., the output signals of the MVDR and the LCMV beamformers, either using reverberant or anechoic RTF vectors.
- unprocessed microphone signals, i.e., the reference microphone signal at the side of speaker 1 as the reference signal for speaker 1 and the reference microphone signal at the side of speaker 2 as the reference signal for speaker 2.

3.5.3 *Algorithm implementation details*

3.5.3.1 *DOA estimation algorithm*

For the DOA estimation algorithm (see Section 3.3.1), the SVM classifiers were trained using simulated noisy speech signals, generated by convolving clean speech signals from the TIMIT database with anechoic binaural room impulse responses (BRIRs) from [110] and adding diffuse speech-shaped noise at SNRs of -20 dB to 20 dB in steps of 10 dB. The GCC-PHAT features were calculated using a frame length of 10 ms with an overlap of 5 ms. The smoothed SPP $\overline{p}_\theta$ in (3.7) was initialized with $p_\theta\,[1]$ and recursively smoothed using a corresponding time constant of 1 s. The DOAs of speaker 1 and 2 were then determined as two DOAs between -90° and 90° (in steps of 5°) with the largest smoothed SPP such that $\hat{\theta}_1 \leq \hat{\theta}_2$.

3.5.3.2 *RTF vector estimation algorithm*

The RTF estimation algorithm (see Section 3.3.2) was implemented using a weighted overlap-add (WOLA) framework with different STFT frame lengths, i.e., FL = 512, 1024, 2048, 4096, 8192 samples, and an overlap of 50% between successive frames. The microphone covariance matrix in (3.9) was initialized using the cylindrically isotropic noise assumption and recursively smoothed using a corresponding time constant of 50 ms. The estimated RTF vectors $\hat{\mathbf{a}}_{1,\{L,R\}}\,(k,l)$ in (3.13) corresponding to speaker 1 were initialized with the anechoic RTF vectors corresponding to the estimated DOA of speaker 1 at $l = 1$. Similarly, the estimated RTF vectors $\hat{\mathbf{a}}_{2,\{L,R\}}\,(k,l)$ in (3.14) corresponding to speaker 2 were initialized with the anechoic RTF vectors corresponding to the estimated DOA of speaker 2 at $l = 1$. The estimated RTF vectors in (3.13) and (3.14) were recursively smoothed using a corresponding time constant of 100 ms.

As proposed in [207], the oracle reverberant RTF vectors $\bar{\mathbf{a}}_t$ and $\bar{\mathbf{a}}_i$ corresponding to the target and the interfering speaker were calculated as the normalized principal eigenvector of the (oracle) target and interference covariance matrix, respectively. These covariance matrices were constructed using white noise convolved with the reverberant BRIRs from [110] corresponding to the target and the interfering speaker.

Similarly, the anechoic RTF vectors $\mathbf{a}(\theta)$ for angle θ was calculated as the normalized principal eigenvector of the (oracle) covariance matrix for angle θ, constructed using white noise convolved with the anechoic BRIRs from [110] for angle θ.

3.5.3.3 *Beamforming*

All considered beamformers were implemented using WOLA framework with a default STFT frame length FL = 512 when using anechoic RTF vectors and a default STFT frame length FL = 8192 when using reverberant RTF vectors, with an overlap of 50% between successive frames. To investigate the impact of the STFT frame length on the AAD performance and the speech enhancement performance when using reverberant RTF vectors, we will also consider FL = 1024, 2048, 4096.

To investigate the difference between using reverberant or anechoic RTF vectors and to investigate the impact of RTF, DOA and AAD estimation errors on the speech enhancement performance of the complete proposed system, we will consider the following combinations:

- ORTF–OAAD, with oracle reverberant RTF vectors and oracle AAD;
- ORTF–EAAD, with oracle reverberant RTF vectors and estimated AAD;
- ODOA–OAAD, with anechoic RTF vectors using the oracle DOAs and oracle AAD;
- ODOA–EAAD, with anechoic RTF vectors using the oracle DOAs and estimated AAD;
- ERTF–EAAD, with estimated reverberant RTF vectors using the estimated DOAs and estimated AAD;
- EDOA–EAAD, with anechoic RTF vectors using the estimated DOAs and estimated AAD.

Please note that for the complete system we will either use the binaural MVDR beamformer (to generate the binaural output signals) together with MVDR beamformers (to generate the reference signals), or the binaural LCMV beamformer (to generate the binaural output signals) together with LCMV beamformers (to generate the reference signals). To investigate the impact on the speech enhancement performance as well as the binaural cue preservation of the interference suppression factor δ used in the binaural LCMV beamformer (see Section 3.4.3), we will consider several values for the interference suppression factor, i.e., $\delta = 0,\ 0.1,\ 0.2$. In addition, to compare the proposed cognitive-driven beamformer with a frequently used beamformer in hearing aids, we will also consider the forward-steered binaural MVDR beamformer (FS–BMVDR), i.e., assuming that the attended speaker is located in the frontal direction, corresponding to using anechoic RTF vectors and fixed DOA $\hat{\theta}_a = 0^\circ$ in (3.37) and (3.38).

3.5.4 *Performance measure*

The performance of the MVDR and LCMV beamformers for generating reference signals is evaluated in terms of the signal-to-interference-plus-noise ratio improve-

ment (ΔSINR). The input SINRs in the reference microphones for the beamformer corresponding to speaker 1 (left hearing aid) for the beamformer corresponding to speaker 2 (right hearing aid) are defined as

$$\mathrm{SINR}_{in,1} = 10\log_{10}\frac{\varepsilon\left\{|x_{1,L,1}|^2\right\}}{\varepsilon\left\{|x_{2,L,1}+v_{L,1}|^2\right\}}, \tag{3.53}$$

$$\mathrm{SINR}_{in,2} = 10\log_{10}\frac{\varepsilon\left\{|x_{2,R,1}|^2\right\}}{\varepsilon\left\{|x_{1,R,1}+v_{R,1}|^2\right\}}. \tag{3.54}$$

The output SINRs for the beamformer corresponding to speaker 1 and the beamformer corresponding to speaker 2 are defined as

$$\mathrm{SINR}_{out,1} = 10\log_{10}\frac{\varepsilon\left\{|z_{t,1}|^2\right\}}{\varepsilon\left\{|z_{i,1}+z_{v,1}|^2\right\}}, \tag{3.55}$$

$$\mathrm{SINR}_{out,2} = 10\log_{10}\frac{\varepsilon\left\{|z_{t,2}|^2\right\}}{\varepsilon\left\{|z_{i,2}+z_{v,2}|^2\right\}}, \tag{3.56}$$

with all output signal components defined in (3.23) and (3.24). The average SINR improvement for both speakers is defined as

$$\Delta\mathrm{SINR} = \frac{\Delta\mathrm{SINR}_1 + \Delta\mathrm{SINR}_2}{2}, \tag{3.57}$$

with

$$\Delta\mathrm{SINR}_1 = \mathrm{SINR}_{out,1} - \mathrm{SINR}_{in,1}, \tag{3.58}$$

$$\Delta\mathrm{SINR}_2 = \mathrm{SINR}_{out,2} - \mathrm{SINR}_{in,2}. \tag{3.59}$$

To evaluate the AAD performance, for each trial the correlation coefficients corresponding to the (oracle) attended speaker ρ_a and the (oracle) unattended speaker ρ_u are computed. A trial is considered to be correctly decoded if $\rho_a > \rho_u$. The AAD performance is then computed by averaging the percentage of correctly decoded trials over all considered trials and all participants.

The speech enhancement performance of the binaural MVDR and LCMV beamformers is evaluated in terms of the binaural signal-to-interference-plus-noise ratio improvement (ΔBSINR). The binaural input SINR is defined as

$$\mathrm{BSINR}_{in} = 10\log_{10}\frac{\varepsilon\left\{|x_{a,L,1}|^2\right\} + \varepsilon\left\{|x_{a,R,1}|^2\right\}}{\varepsilon\left\{|x_{u,L,1}+v_{L,1}|^2\right\} + \varepsilon\left\{|x_{u,R,1}+v_{R,1}|^2\right\}}, \tag{3.60}$$

where $x_{a,L,1}$ and $x_{a,R,1}$ denote the (oracle) attended speech components in the reference microphone signals, and $x_{u,L,1}$ and $x_{u,R,1}$ denote the (oracle) unattended speech components in the reference microphone signals. The binaural output SINR is defined as

$$\mathrm{BSINR}_{out} = 10\log_{10}\frac{\varepsilon\left\{|z_{a,L}|^2\right\}+\varepsilon\left\{|z_{a,R}|^2\right\}}{\varepsilon\left\{|z_{u,L}+z_{v,L}|^2\right\}+\varepsilon\left\{|z_{u,R}+z_{v,R}|^2\right\}}, \tag{3.61}$$

with all output signal components defined in (3.49) and (3.50). The BSINR improvement is defined as

$$\Delta\mathrm{BSINR} = \mathrm{BSINR}_{out} - \mathrm{BSINR}_{in}. \tag{3.62}$$

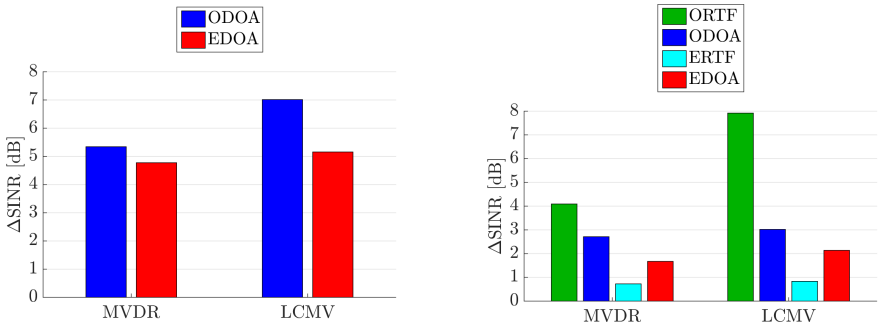
To evaluate the binaural cue preservation of the unattended speaker at the output of the binaural beamformers, we calculate the ILD and ITD errors, averaged over all frequencies, using the binaural auditory model proposed in [208].

3.6 Results and discussion

In this section, we evaluate the AAD performance and the speech enhancement performance of the proposed cognitive-driven binaural beamforming system using the experimental setup discussed in the previous section. In Section 3.6.1 we evaluate the SINR improvement of the beamformers used to generate reference signals for decoding, where we also investigate the difference between using reverberant or anechoic RTF vectors and the impact of RTF and DOA estimation errors. In Section 3.6.2 we evaluate the decoding performance using these reference signals. Finally, in Section 3.6.3 we evaluate the speech enhancement performance of the binaural beamformers, where in addition to RTF and DOA estimation errors we also investigate the impact of AAD errors.

3.6.1 *Performance of the beamformers for generating reference signals*

For the anechoic and the reverberant condition, Fig. 3.3 depicts the average SINR improvement in (3.57) of the MVDR and LCMV beamformers used to generate reference signals. When using oracle (anechoic or reverberant) RTF vectors, i.e., ODOA for the anechoic condition and ORTF for the reverberant condition, it can be observed that an SINR improvement of about 4 - 5 dB is obtained by the MVDR beamformers, while a larger SINR improvement of about 7 - 8 dB is obtained by the LCMV beamformers. The larger SINR improvement obtained by the LCMV beamformers can be explained by the interference suppression constraint in (3.17), which leads to a larger suppression of the interfering speaker (and a similar noise reduction) compared to the MVDR beamformers [9,205]. When using anechoic RTF vectors (ODOA) instead of reverberant RTF vectors (ORTF) in the reverberant condition, it can be observed for both beamformers that the SINR improvement



(a) anechoic condition (b) reverberant condition

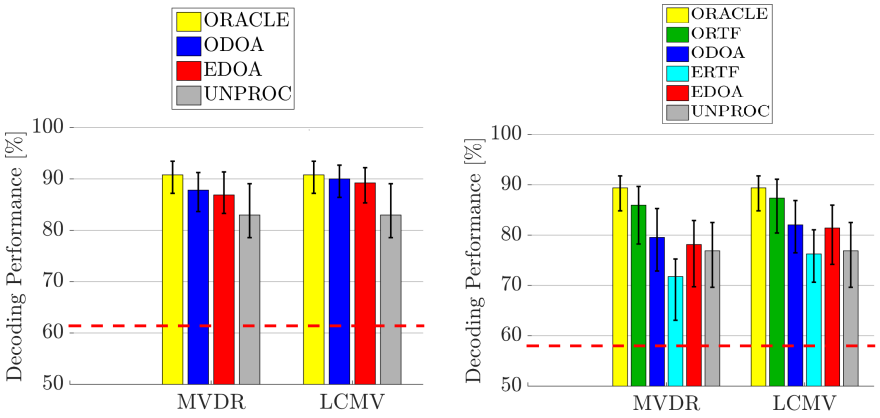
Fig. 3.3: Average SINR improvement of the MVDR and LCMV beamformers for (a) the anechoic condition and (b) the reverberant condition using oracle RTFs, oracle DOAs, estimated RTFs and estimated DOAs.

substantially decreases. The decrease is larger for the LCMV beamformers compared to the MVDR beamformers, mainly due to the fact that the interfering speaker is suppressed less than when using imperfect RTF vectors (i.e., anechoic RTF vectors in the reverberant condition). Nevertheless, the SINR improvement obtained by the LCMV beamformers is still larger than the MVDR beamformer. When using estimated RTF vectors (ERTF) in the reverberant condition, it can be observed that the SINR improvement decreases rather considerably by 3.2 - 7 dB compared to using oracle RTF vectors (ORTF). This can be explained by the fact that the reverberant RTF vector estimation method presented in Section 3.3.2 is not able to accurately estimate the RTF vectors for both speakers. However, when using estimated DOAs (EDOA), the SINR improvement for both beamformers and for both acoustic conditions decreases only by 0.9 - 1.1 dB compared to using oracle DOAs (ODOA).

3.6.2 *Auditory attention decoding performance*

For the anechoic and the reverberant condition, Fig. 3.4 depicts the average decoding performance when using either the oracle anechoic signals, the MVDR output signals, the LCMV output signals, or the unprocessed microphone signals as reference signals for decoding. For both acoustic conditions, it can be observed that the largest decoding performance is obtained when using the oracle anechoic signals ($> 89\%$) and the worst decoding performance is obtained when using either the unprocessed microphone signals ($> 77\%$) or the beamformer output signals with estimated RTF vectors (ERTF) ($> 71\%$).

When using oracle RTF vectors, i.e., ODOA for the anechoic condition and ORTF for the reverberant condition, the average decoding performance for both beamformers is substantially larger than when using the unprocessed microphone signals. When using anechoic RTF vectors (ODOA) instead of reverberant RTF vectors



(a) anechoic condition (b) reverberant condition

Fig. 3.4: Average decoding performance for (a) the anechoic condition and (b) the reverberant condition when using the oracle anechoic signals, the MVDR output signals, the LCMV output signals, and the unprocessed microphone signals as reference signals for decoding. The red dashed-line represents the upper boundary of the confidence interval corresponding to chance level based on a binomial test at the 5% significance level. The error bars represent the bootstrap confidence interval at the 5% significance level.

(ORTF) in the reverberant condition, it can be observed that the decoding performance substantially decreases but is still larger than the decoding performance using the unprocessed microphone signals. When using estimated RTF vectors (ERTF) in the reverberant condition, the decoding performance is even lower than when using the unprocessed microphone signals, showing that for the considered acoustic setup the AAD performance is sensitive to RTF estimation errors. However, when using estimated DOAs (EDOA), it can be observed for both beamformers and for both acoustic conditions that the decoding performance is larger than when using the unprocessed microphone signals. The decoding performance for the LCMV beamformers ($> 82\%$) is larger than for the MVDR beamformers ($> 77\%$), which can be explained by the larger SINR improvement of the LCMV beamformers and especially the larger interference suppression compared to the MVDR beamformers (see Fig. 3.3). This is in accordance with the experimental results in [26, 32], where it has been shown that jointly suppressing interference and background noise is of great importance for reference signal generation. In addition, it can be observed for both beamformers and both acoustic conditions that the decoding performance using estimated DOAs (EDOA) is very similar to using oracle DOAs (ODOA).

To investigate the impact of the STFT frame length, Fig. 3.5 depicts the average decoding performance in the reverberant condition for several STFT frame lengths when using either the MVDR output signals or the LCMV output signals with different RTF vectors. These results show that the decoding performance is very similar for all considered STFT frame lengths.

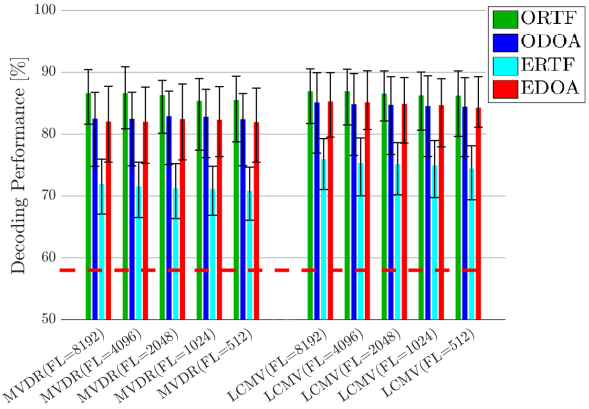


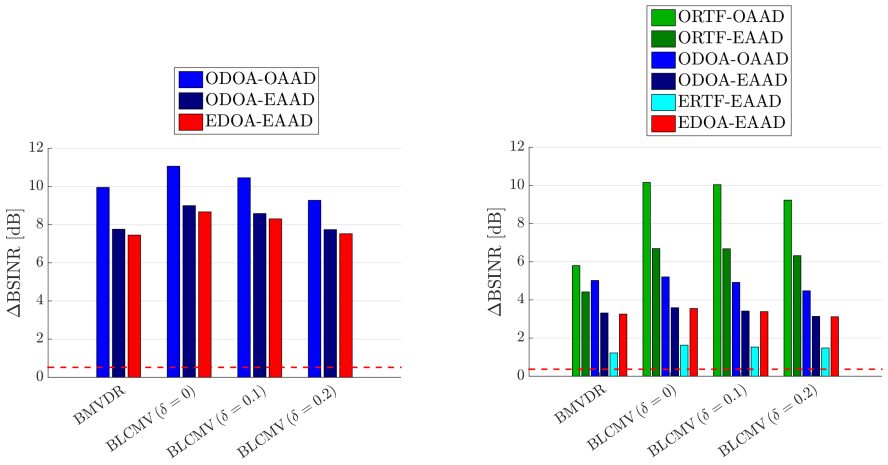
Fig. 3.5: Average decoding performance using different STFT frame lengths in the reverberant condition. The red dashed-line represents the upper boundary of the confidence interval corresponding to chance level based on a binomial test at the 5% significance level. The error bars represent the bootstrap confidence interval at the 5% significance level.

3.6.3 *Binaural speech enhancement performance*

For the anechoic and the reverberant condition, Fig. 3.6 depicts the binaural SINR improvement of the complete proposed system using either the binaural MVDR beamformer or the binaural LCMV beamformer (for several values of the interference suppression factor δ). In addition, this figure depicts the binaural SINR improvement of the forward-steered binaural MVDR beamformer.

When using both oracle AAD as well as oracle RTF vectors, i.e., ODOA–OAAD in the anechoic condition and ORTF–OAAD in the reverberant condition, it can be observed that the binaural MVDR beamformer yields a binaural SINR improvement of 9.5 dB (anechoic condition) and 5.8 dB (reverberant condition), while the binaural LCMV beamformer yields a binaural SINR improvement of 9.3 - 11.0 dB (anechoic condition) and 9.2 - 10.2 dB (reverberant condition). When using anechoic RTF vectors (ODOA–OAAD) instead of reverberant RTF vectors (ORTF–OAAD) in the reverberant condition, it can be observed that the binaural SINR improvement of both beamformers substantially decreases, i.e., 4.4 dB for the binaural MVDR beamformer and 6.3 - 6.7 dB for the binaural LCMV beamformer.

When using oracle RTF vectors and estimated AAD, i.e., ODOA–EAAD in the anechoic condition and ORTF–EAAD or ODOA–EAAD in the reverberant condition, it can be observed that the binaural SINR improvement decreases for both beamformers compared to using oracle AAD. The decrease is especially significant for the LCMV beamformer using oracle reverberant RTF vectors in the reverberant condition. When using estimated RTF vectors and estimated AAD (ERTF–EAAD) in the reverberant condition, it can be observed that the binaural SINR improvement significantly decreases for both beamformers compared to oracle RTF vec-

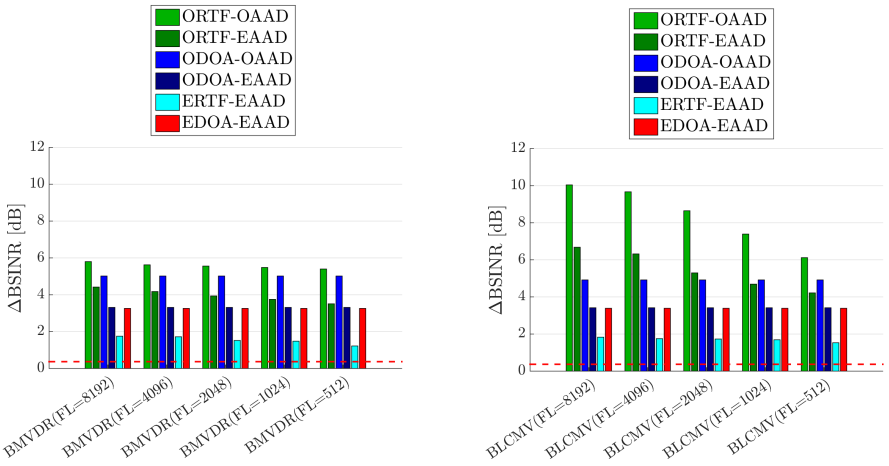


(a) anechoic condition (b) reverberant condition

Fig. 3.6: Binaural SINR improvement of the proposed system when using the binaural MVDR beamformer (BMVDR) and the binaural LCMV beamformer (BLCMV) for several values of the interference suppression factor δ for (a) the anechoic condition and (b) the reverberant condition. The red dashed-line represents the binaural SINR improvement of the forward-steered binaural MVDR beamformer (FS–BMVDR).

tors (ORTF–EAAD). However, when using estimated DOAs and estimated AAD (EDOA–EAAD), a very similar binaural SINR improvement is obtained for both beamformers and both acoustic conditions compared to using oracle DOAs. These results clearly show that for the reverberant condition the practically implementable EDOA–EAAD system (using estimated DOAs and anechoic RTF vectors) outperforms the practically implementable ERTF–EAAD systems (using estimated reverberant RTF vectors).

To investigate the impact of the STFT frame length, Fig. 3.7 depicts the binaural SINR improvement in the reverberant condition for different STFT frame lengths when using the binaural MVDR beamformer or the binaural LCMV beamformer with interference suppression factor $\delta = 0.1$. On the one hand, when using reverberant RTF vectors (ORTF–OAAD, ORTF–EAAD, ERTF–EAAD), it can be observed for both beamformers that the binaural SINR improvement decreases for smaller STFT frame lengths. In general, the impact of the STFT frame length is larger for the LCMV beamformer than for the MVDR beamformer, since a larger frame length leads to a larger suppression of the interfering speaker (especially when using oracle reverberant RTF vectors). On the other hand, when using anechoic RTF vectors (ODOA–OAAD, ODOA–EAAD, EDOA–EAAD), the frame length only has a minor impact on the binaural SINR improvement. These results hence show that the practically implementable system using estimated AAD together with estimated DOAs yields a large binaural SINR improvement even when rather using shorter

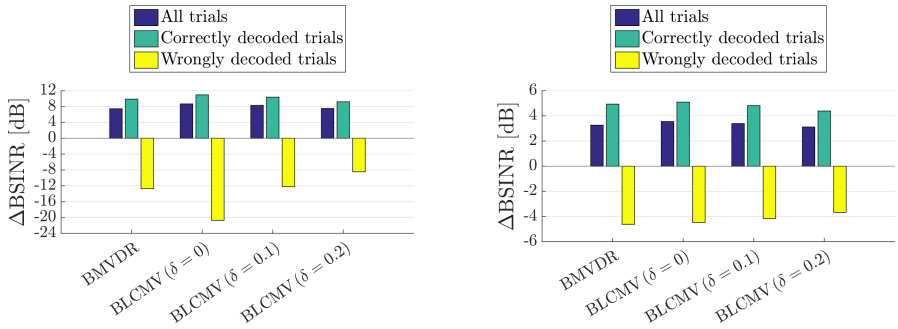


(a) binaural MVDR beamformer (b) binaural LCMV beamformer

Fig. 3.7: Binaural SINR improvement of the proposed system using different STFT frame lengths in the reverberant condition for (a) the binaural MVDR beamformer (BMVDR) and (b) the binaural LCMV beamformer (BLCMV). The red dashed-line represents the binaural SINR improvement of the forward-steered binaural MVDR beamformer (FS–BMVDR).

STFT frames. In a practical implementation, the latency of the proposed system using estimated AAD and estimated DOAs consists of three parts: AAD estimation, DOA estimation and STFT-domain processing. The latency caused by AAD estimation in (3.31) using 30 second trials corresponds to 30 s. The latency caused by DOA estimation in (3.7) using a frame length of 10 ms with an overlap of 5 ms and a time constant of 1 s corresponds to 1.115 s. The latency caused by STFT-domain processing in (3.37), (3.38), (3.43) and (3.44) using an STFT frame length of 512 samples with an overlap of 50% between successive frames corresponds to 64 ms. It should be noted that the latency caused by AAD and DOA estimation only affects the startup time and does not affect the processing latency of the binaural signals, which is only determined by the STFT-domain processing.

To further investigate the impact of AAD errors on the binaural SINR improvement when using estimated AAD and estimated DOAs (EDOA–EAAD), Fig. 3.8 depicts the binaural SINR improvement averaged over all trials (as in Fig. 3.7) and the binaural SINR improvement averaged only over correctly decoded and wrongly decoded trials. When trials are wrongly decoded, the unattended speaker is wrongly enhanced by the binaural MVDR and LCMV beamformer and in addition the attended speaker is wrongly suppressed by the LCMV beamformer, such that the binaural SINR improvement averaged over wrongly decoded trials is negative for both beamformers. For the binaural LCMV beamformer with $\delta = 0.2$, the binaural SINR improvement is less prone to wrongly decoded trials compared to the binaural LCMV beamformer with a smaller δ. Nevertheless, the binaural LCMV beamformer



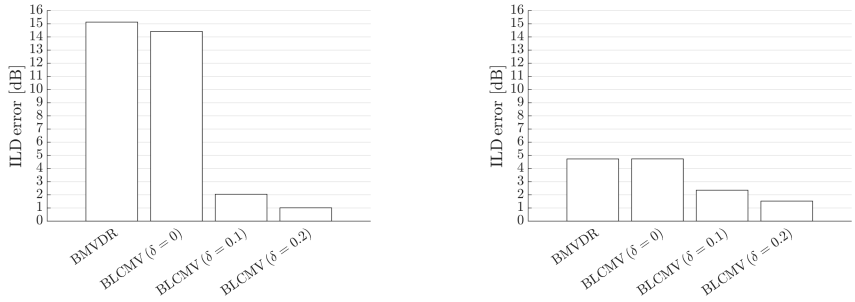
(a) anechoic condition (b) reverberant condition

Fig. 3.8: Binaural SINR improvement averaged over all trials, all correctly decoded trials and all wrongly decoded trials obtained when using estimated DOAs and estimated AAD (EDOA–EAAD) for (a) the anechoic condition and (b) the reverberant condition.

with $\delta = 0$ or $\delta = 0.1$ yields the largest binaural SINR improvement averaged over all (correctly and wrongly decoded) trials, i.e., 8.3 - 8.7 dB (anechoic condition) and 3.4 - 3.6 dB (reverberant condition). This is larger than the binaural SINR improvement of the cognitive-driven binaural MVDR beamformer, i.e., 7.4 dB (anechoic condition) and 3.2 dB (reverberant condition), and significantly larger than the binaural SINR improvement of the forward-steered binaural MVDR beamformer, i.e., 0.3 dB (anechoic condition) and 0.5 dB (reverberant condition).

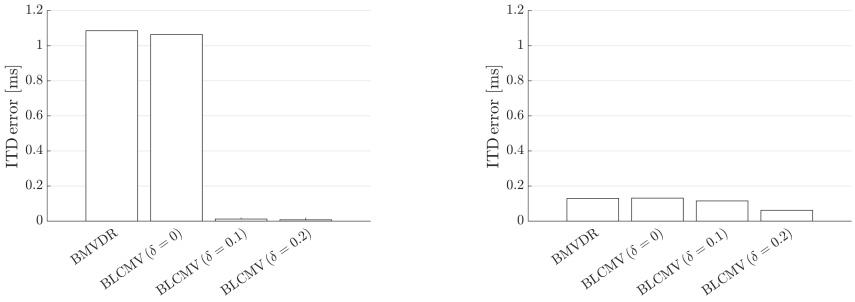
Finally, we evaluate the binaural cue preservation of the unattended speaker, i.e., how well the impression of the acoustic scene is preserved. For the anechoic and the reverberant condition, Fig. 3.9 and 3.10 present the ILD and ITD errors of the unattended speaker (averaged only over correctly decoded trials) when using estimated DOAs and estimated AAD (EDOA–EAAD). It can be observed that the binaural MVDR beamformer and the binaural LCMV beamformer with $\delta = 0$ yield large ILD and ITD errors, while the binaural LCMV with $\delta > 0$ yields a better binaural cue preservation for both acoustic conditions. The better binaural cue preservation obtained by the binaural LCMV beamformer can be explained by considering the role of the interference suppression constraint in the optimization problem of the binaural LCMV beamformer (see Section 3.4.3). The interference suppression factor $\delta > 0$ allows the binaural LCMV beamformer to preserve the binaural cues of the unattended speaker in addition to the binaural cues of the attended speaker, contrary to the binaural MVDR beamformer [91, 122].

Table 3.2 compares the performance of all considered beamformers in terms of binaural SINR improvement, binaural cue preservation, and the impact of AAD, RTF vector and DOA estimation errors on the performance.



(a) anechoic condition (b) reverberant condition

Fig. 3.9: ILD errors of the unattended speaker averaged over correctly decoded trials when using estimated DOAs and estimated AAD (EDOA–EAAD) for (a) the anechoic condition and (b) the reverberant condition.



(c) anechoic condition (d) reverberant condition

Fig. 3.10: ITD errors of the unattended speaker averaged over correctly decoded trials when using estimated DOAs and estimated AAD (EDOA–EAAD) for (a) the anechoic condition and (b) the reverberant condition.

3.7 Conclusion

In this paper, we proposed a binaural speech enhancement system which cognitively steers the binaural MVDR and the binaural LCMV beamformer based on AAD and estimated DOA-based anechoic or reverberant RTF vectors. Based on these RTF vectors, two MVDR or LCMV beamformers generate reference signals for auditory attention decoding. Using the envelopes of these reference signals and the EEG recordings, in the AAD step the DOAs of the attended and the unattended speaker are identified to steer the binaural MVDR or LCMV beamformer. The experimental results showed that for a two-speaker scenario in diffuse babble noise the proposed system using anechoic DOA-based RTF vectors significantly improves the binaural SINR for the anechoic condition as well as for the reverberant condition compared to a fixed forward-steered binaural MVDR beamformer. In particular, the cognitive-

Table 3.2: Comparison of the proposed cognitive-driven speech enhancement system using either the binaural MVDR beamformer or the binaural LCMV beamformer and the forward-steered binaural MVDR beamformer

	Binaural MVDR	**Binaural LCMV with $\delta = 0$**	**Binaural LCMV with $\delta = 0.1$**	**Binaural LCMV with $\delta = 0.2$**	**Forward-steered Binaural MVDR**
Binaural SINR improvement	Medium	Very large	Large	Medium	Low
Binaural cues of unattended speaker	Not preserved	Not preserved	Preserved	Preserved	Not preserved
Impact of DOA errors	Low	Low	Low	Low	DOA–independent
Impact of RTF errors	High	High	High	High	RTF–independent
Impact of AAD errors	Medium	High	Medium	Medium	AAD–independent

driven binaural LCMV beamformer with $\delta = 0.1$ is able to both improve the binaural SINR as well as preserve the binaural cues of both the attended and the unattended speaker. Moreover, the results show that for the considered experimental setup the proposed system using estimated DOA-based anechoic RTF vectors yields a larger binaural SINR improvement for the reverberant condition compared to using estimated DOA-based reverberant RTF vectors. Furthermore, the results show that the STFT frame length only has a minor impact on the binaural SINR improvement when using estimated DOA-based anechoic RTF vectors.

While the application of the proposed cognitive-driven binaural speech enhancement system has been limited to acoustic scenarios with two competing speakers in this paper, in [209] it has been shown that AAD is feasible for an acoustic scenario with four competing speakers when using perfectly separated clean speech signals for decoding. Future work could therefore investigate the performance of (an extension of) the proposed cognitive-driven binaural speech enhancement system for acoustic scenarios with more than two competing speakers.

4

COGNITIVE-DRIVEN CONVOLUTIONAL BEAMFORMING USING EEG-BASED AUDITORY ATTENTION DECODING

The performance of speech enhancement algorithms in a multi-speaker scenario depends on correctly identifying the target speaker to be enhanced. Auditory attention decoding (AAD) methods allow to identify the target speaker which the listener is attending to from single-trial EEG recordings. Aiming at enhancing the target speaker and suppressing interfering speakers, reverberation and ambient noise, in this paper we propose a cognitive-driven multi-microphone speech enhancement system, which combines a neural-network-based mask estimator, weighted minimum power distortionless response convolutional beamformers and AAD. To control the suppression of the interfering speaker, we also propose an extension incorporating an interference suppression constraint. The experimental results show that the proposed system outperforms the state-of-the-art cognitive-driven speech enhancement systems in challenging reverberant and noisy conditions.

4.1 Introduction

In a multi-speaker scenario the performance of many speech enhancement algorithms depends on correctly identifying the target speaker to be enhanced. Recent advances in electroencephalography (EEG) have shown that it is possible to identify the target speaker which the listener is attending to using single-trial EEG-based auditory attention decoding (AAD) methods [16,26,128,130]. However, many AAD methods rely on the unrealistic assumption that the clean speech signals of the speakers are available as reference signals for decoding. In real-world conditions, obviously only the microphone signals, which consist of a mixture of the speakers, including reverberation and background noise, are available.

Aiming at incorporating AAD in speech enhancement, several algorithms have recently been proposed to generate appropriate reference signals for decoding from the microphone signals [27, 30, 31, 167]. Most cognitive-driven speech enhancement algorithms generate reference signals by separating the speakers from the mixture received at the microphones either using time-domain neural networks [167], multi-channel Wiener filters [27] or minimum variance distortionless response (MVDR)

beamformers [31]. Using AAD, one of the reference signals is then selected as the enhanced attended speaker. More recently, aiming at controlling the suppression of the interfering speaker, which is important when intending to switch attention between speakers, a cognitive-driven beamforming system using linearly constrained minimum variance (LCMV) beamformers has been proposed [30, 31].

While most aforementioned cognitive-driven speech enhancement systems are able to suppress the interfering speakers and background noise, they may not be able to suppress (late) reverberation, which is known to have a detrimental effect on speech quality and intelligibility [39]. In this paper we propose a cognitive-driven convolutional beamforming system aiming at enhancing the attended speaker and jointly suppressing the interfering speakers, reverberation and background noise.

The proposed system is depicted in Fig. 4.1 for a scenario with two speakers. First, time-frequency masks of both speakers are estimated from the noisy and reverberant microphone signals using a speaker-independent speech separation neural network. Then, two beamformers are designed to generate reference signals for AAD by enhancing the speech signal of each speaker based on the estimated masks. The AAD method then selects one of the reference signals as the enhanced attended speech signal. For the beamformers we propose to use a recently proposed weighted minimum power distortionless response (wMPDR) convolutional beamformer as it optimally combines dereverberation, noise suppression and interfering speaker suppression [100]. While suppressing the interfering speaker is desired to improve speech intelligibility, keeping the interfering speaker audible is also important to allow the listener to switch attention between speakers. Therefore, we also propose an extension of the wMPDR convolutional beamformer incorporating an interference suppression constraint, referred to as a weighted linearly constrained minimum power (wLCMP) convolutional beamformer, which allows to control the level of suppression of the interfering speaker.

We experimentally compare our proposed method with state-of-the-art cognitive-driven systems based on conventional MPDR, LCMP, MVDR and LCMV beamformers, which are steered based on estimated masks or estimated DOAs. The results show that the proposed system outperforms state-of-the-art cognitive-driven systems for dealing with noisy and reverberant speech mixtures and reveal potential future research directions.

4.2 Cognitive-driven convolutional beamformer

4.2.1 *Signal model*

We consider an acoustic scenario comprising I competing speakers[1] with the clean signals denoted as $s_i[n]$, $i = 1 \ \ldots \ I$ where n is the discrete time index. We consider

1 It should be noted that we provide a general description of the algorithms for I speakers, but limit our experiments in Section 4.4 to two speakers.

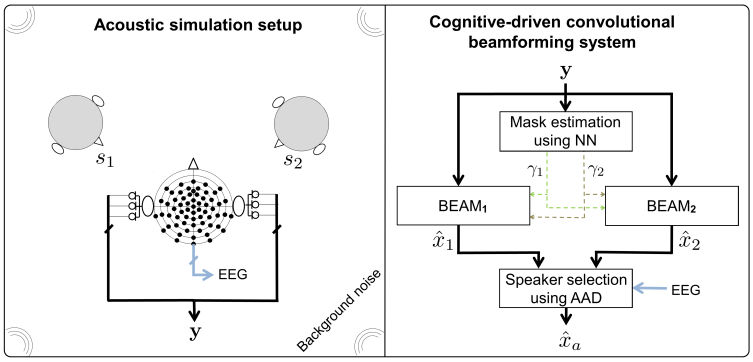


Fig. 4.1: Acoustic simulation setup and block diagram of the proposed cognitive-driven convolutional beamforming system.

a binaural hearing aid setup with M microphones. The m-th microphone signal $y_m[n]$ can be decomposed as

$$y_m[n] = \sum_{i=1}^{I} x_{i,m}[n] + v_m[n], \qquad m = 1 \ \ldots \ M, \tag{4.1}$$

where $x_{i,m}[n]$ denotes the reverberant speech component in the m-th microphone signal corresponding to speaker i and $v_m[n]$ denotes the background noise component. The reverberant speech components $x_{i,m}[n]$ consist of an anechoic speech component $x_{i,m}^{an}[n]$ (encompassing the head filtering effect), an early reverberation component arriving typically in the order of tens of milliseconds, and a late reverberation component. While early reverberation can be beneficial for speech intelligibility, late reverberation is known to have a detrimental effect on speech quality and intelligibility [39].

In the short-time Fourier Transform (STFT) domain, the M-dimensional stacked vector of all microphone signals is given by

$$\mathbf{y}_{k,f} = [Y_{1,k,f} \ \ldots \ Y_{M,k,f}]^T \in \mathbb{C}^{M\times 1}, \tag{4.2}$$

where $Y_{m,k,f}$ denotes the STFT coefficient of $y_m[n]$, and $k = 1 \ \ldots \ K$ and $f = 1 \ \ldots \ F$ are the frame index and the frequency index, respectively.

4.2.2 *Mask estimation*

The first component of our proposed system is a separation neural network that estimates time-frequency ideal ratio masks corresponding to each speaker from the reverberant and noisy microphone signals. These masks will be used for beamforming and to generate reference signals for AAD (see Section 4.2.3).

Several neural network-based speech separation approaches have been proposed, both in frequency-domain and in time-domain [77, 83]. In this paper we use a BLSTM-based frequency-domain approach [83] since it trains faster than time-domain approaches such as [77], allowing a faster experimental turnover.

The separation neural network takes the STFT coefficients of the m-th microphone signal as input features and generates real-valued time-frequency masks, i.e.,

$$[\mathbf{\Gamma}_{1,m} \ \dots \ \mathbf{\Gamma}_{I+1,m}] = h(\mathbf{Y}_m), \tag{4.3}$$

where the matrix $\mathbf{Y}_m \in \mathbb{C}^{K\times F}$ contains all STFT coefficients of the m-th microphone signal, $h(\cdot)$ is the separation neural network, and the matrix $\mathbf{\Gamma}_{i,m} \in \mathbb{R}^{K\times F}$ for $i = 1 \ \dots \ I$, contains the estimated time-frequency masks for speaker i. In addition to the time-frequency masks for the speakers, the network also generates a time-frequency mask for the background noise, i.e., $\mathbf{\Gamma}_{I+1,m}$.

The separation neural network is trained using permutation invariant training (PIT) [83] with a scale-dependent SNR loss in the time-domain [210]. However, at test time the masks have speaker permutation ambiguity, i.e., it is not known which mask corresponds to which speaker. In addition, the separation neural network in (4.3) operates on each microphone signal independently, which typically causes speaker permutation ambiguities across the microphones. To resolve this ambiguity, we align the masks obtained for each microphone based on the least-squares error. We then average the masks across the microphones to obtain one mask for each speaker, i.e. $\bar{\mathbf{\Gamma}}_i \in \mathbb{R}^{K\times F}$. The averaged mask $\bar{\mathbf{\Gamma}}_i$ contains the masks $\gamma_{i,k,f}$ of the i-th speaker for all times frames and frequencies.

4.2.3 *Reference signal generation using beamformers*

Based on the estimated masks $\bar{\mathbf{\Gamma}}_i$, we design I beamformers to extract each speaker with reduced noise and reverberation from the microphone signals (see BEAM_1 and BEAM_2 in Fig. 4.1). The output signals $z_{i,k,f}$ of the beamformers are then transformed to the time-domain as $\hat{x}_i[n] = \text{ISTFT}(z_{i,k,f})$, where ISTFT denotes the inverse short-time Fourier transform. These time-domain output signals $\hat{x}_i[n]$ will be used as reference signals for AAD.

In this paper we investigate different types of beamformers for generating reference signals, i.e., wMPDR and wLCMP convolutional beamformers, and conventional MPDR and LCMP beamformers, which will be described in detail in Section 4.3.

4.2.4 *Speaker selection using AAD*

Based on the reference signals $\hat{x}_i[n]$ generated by the beamformers, the speaker which the listener is attending to is then selected using the EEG-based auditory attention decoding method proposed in [16]. First, an estimate of the envelope of the attended speech signal $\hat{e}_a[l]$, with l the sub-sampled time index, is reconstructed from the EEG signals using a trained spatio-temporal filter. Then, the correlation

between the reconstructed envelope $\hat{e}_a[l]$ and the envelopes $\hat{e}_i[l]$ of the reference signals $\hat{x}_i[n]$ is computed, i.e.,

$$\rho_i = \rho\left(\hat{e}_i[l], \quad \hat{e}_a[l]\right), \qquad i = 1 \ \ldots \ I, \tag{4.4}$$

where $\rho(\cdot)$ is the Pearson correlation. Finally, the attended speech signal $\hat{x}_a[n]$ is selected as the reference signal yielding the maximum correlation with the reconstructed envelope, i.e.,

$$\hat{x}_a[n] = \hat{x}_{\bar{i}}[n], \qquad \bar{i} = \underset{i}{\operatorname{argmax}} \ \rho_i. \tag{4.5}$$

4.3 Beamforming

In this section, we review the wMPDR convolutional beamformer [98], present the proposed wLCMP convolutional beamformer, and compare them with the conventional MPDR and LCMP beamformers. Since the beamformer operates for each frequency independently, the frequency index f will be omitted in this section for notational conciseness.

4.3.1 *Weighted MPDR convolutional beamformer*

The wMPDR convolutional beamformer in [98] aims at 1) suppressing the noise component while preserving the target speech component in one of the microphone signals and 2) suppressing the late reverberation component while preserving the early reverberation component corresponding to the target speaker (i.e., dereverberation). The output signal z_k of a convolutional beamformer is defined as

$$z_k = \bar{\mathbf{w}}^H \bar{\mathbf{y}}_k = \mathbf{w}_0^H \mathbf{y}_k + \sum_{\tau=b}^{L_w-1} \mathbf{w}_\tau^H \mathbf{y}_{k-\tau}, \tag{4.6}$$

where $\bar{\mathbf{w}} = \left[\mathbf{w}_0^T \ \mathbf{w}_b^T \ldots \ \mathbf{w}_{L_w-1}^T\right]^T \in \mathbb{C}^{M(L_w-b+1)\times 1}$, $\bar{\mathbf{y}}_k = \left[\mathbf{y}_k^T \ \tilde{\mathbf{y}}_k^T\right]^T \in \mathbb{C}^{M(L_w-b+1)\times 1}$, $\tilde{\mathbf{y}}_k$ consists of the observation from b frames in the past until $L_w - 1$ frames in the past, i.e., $\tilde{\mathbf{y}}_k = \left[\mathbf{y}_{k-b}^T \cdots \ \mathbf{y}_{k-L_w+1}^T\right]^T$, and b and L_w model the frame delay of the start and end time of the late reverberation, respectively.

It has been shown in [100] that the convolutional beamformer $\bar{\mathbf{w}}$ can be factorized into a dereverberation matrix $\mathbf{G} \in \mathbb{C}^{M(L_w-b+1)\times M}$ and a beamforming vector $\mathbf{q} \in \mathbb{C}^{M\times 1}$, i.e., $\bar{\mathbf{w}} = -\mathbf{G}\mathbf{q}$ with $\mathbf{q} = \mathbf{w}_0$. The convolutional beamforming in (4.6) can hence be written as dereverberation filtering followed by beamforming [100], i.e.,

$$\mathbf{d}_k = \underbrace{\mathbf{y}_k - \mathbf{G}^H \bar{\mathbf{y}}_k}_{\text{dereverberation}}, \qquad z_k = \underbrace{\mathbf{q}^H \mathbf{d}_k}_{\text{beamforming}}. \tag{4.7}$$

Assuming that the output of the convolutional beamformer z_k follows a zero mean complex Gaussian distribution with a time-varying variance [98], the wMPDR con-

volutional beamformer is obtained by maximizing an objective function $\mathcal{L}(\bar{\mathbf{w}})$, which is derived based on the maximum-likelihood estimation with a target speaker preservation constraint (distortionless constraint), i.e.,

$$\mathcal{L}(\bar{\mathbf{w}}) \propto \frac{1}{K}\sum_{k=1}^{K}\left(-\ln(\lambda_k) - \frac{|z_k|^2}{\lambda_k}\right), \tag{4.8}$$

where λ_k denotes the time-varying variance of the target speech component (including the early reverberation) and K denotes the number of frames over which the beamformer coefficients are estimated.

This optimization problem can be solved in a alternating fashion, by first assuming λ_k constant and solving for $\bar{\mathbf{w}}$ and then updating λ_k. Assuming λ_k constant, the optimization problem of the wMPDR convolutional beamformer incorporating the target speaker preservation constraint can be written as [98]

$$\max_{\bar{\mathbf{w}}} \; -\bar{\mathbf{w}}^H \bar{\mathbf{R}}_{\bar{\mathbf{y}}} \bar{\mathbf{w}} \quad \text{s.t.} \quad \underbrace{\bar{\mathbf{w}}^H \bar{\mathbf{a}} = 1}_{\text{target}}, \tag{4.9}$$

where $\bar{\mathbf{a}}$ denotes the relative early transfer function (RETF) vector corresponding to the target speaker and $\bar{\mathbf{R}}_{\bar{y}} = \frac{1}{K}\sum_k \frac{\bar{\mathbf{y}}_k \bar{\mathbf{y}}_k^H}{\lambda_k}$. The wMPDR convolutional beamformer solving (4.9) is given by [100]

$$\bar{\mathbf{w}}_{\text{wMPDR}} = -\mathbf{G}\mathbf{q}_{\text{wMPDR}}, \tag{4.10}$$

where

$$\mathbf{G} = \mathbf{R}_{\tilde{y}}^{-1}\mathbf{P}_{\tilde{y}}, \qquad \mathbf{q}_{\text{wMPDR}} = \frac{\mathbf{R}_d^{-1}\bar{\mathbf{a}}}{\bar{\mathbf{a}}^H \mathbf{R}_d^{-1}\bar{\mathbf{a}}}, \tag{4.11}$$

with $\mathbf{R}_{\tilde{y}} = \frac{1}{K}\sum_k \frac{\tilde{\mathbf{y}}_k \tilde{\mathbf{y}}_k^H}{\lambda_k}$, $\mathbf{P}_{\tilde{y}} = \frac{1}{K}\sum_k \frac{\tilde{\mathbf{y}}_k \mathbf{y}_k^H}{\lambda_k}$, $\mathbf{R}_d = \frac{1}{K}\sum_k \frac{\mathbf{d}_k \mathbf{d}_k^H}{\lambda_k}$.

To estimate the RETF vector of the target speaker $\bar{\mathbf{a}}$ in (4.11), we use the masks of the target speaker $\gamma_{t,k}$, assuming the target speaker index is t. The RETF vector is estimated using the covariance whitening method [107], i.e.,

$$\bar{\mathbf{a}} = \mathbf{R}_{\widetilde{t+v}} \text{MaxEig}\left(\mathbf{R}_{\widetilde{t+v}}^{-1}\mathbf{R}_t\right), \tag{4.12}$$

where $\mathbf{R}_t = \frac{\sum_k \gamma_{t,k}\mathbf{d}_k\mathbf{d}_k^H}{\sum_k \gamma_{t,k}}$ is the covariance matrix of the target speaker and $\mathbf{R}_{\widetilde{t+v}} = \frac{\sum_k (1-\gamma_{t,k})\mathbf{d}_k\mathbf{d}_k^H}{\sum_k (1-\gamma_{t,k})}$ is the covariance matrix of all interfering speakers and background noise.

The estimation methods discussed in this section are used to iteratively update the output signal of the wMPDR convolutional beamformer. First, the dereverberation filtering in (4.7) is performed using $\mathbf{G}$ in (4.11). Based on the dereverberated signals $\mathbf{d}_k$ and the estimated masks $\gamma_{t,k}$, the RETF vector of the target speaker $\bar{\mathbf{a}}$ is updated using (4.12) to steer the beamformer $\mathbf{q}_{\text{wMPDR}}$ in (4.11). Using the steered

beamformer, the output signal z_k in (4.7) is obtained. The variance of the target speech component is then updated as $\lambda_k = |z_k|^2$ for the next iteration.

4.3.2 *Weighted LCMP convolutional beamformer*

As an alternative to the wMPDR convolutional beamformer, we propose the wLCMP convolutional beamformer, which allows to control the suppression of the interfering speakers. The wLCMP convolutional beamformer is derived by adding interfering speaker suppression constraints to the optimization problem of the wMPDR convolutional beamformer, i.e.,

$$\max_{\bar{\mathbf{w}}} \; -\bar{\mathbf{w}}^H \bar{\mathbf{R}}_{\bar{\mathbf{y}}} \bar{\mathbf{w}} \quad \text{s.t.} \quad \underbrace{\bar{\mathbf{w}}^H \bar{\mathbf{a}} = 1}_{\text{target}}, \quad \underbrace{\bar{\mathbf{w}}^H \bar{\mathbf{B}} = \boldsymbol{\delta}}_{\text{interference}}, \tag{4.13}$$

where $\bar{\mathbf{B}} = \begin{bmatrix} \bar{\mathbf{b}}_1 & \dots & \bar{\mathbf{b}}_U \end{bmatrix}$ contains the RETF vectors of U interfering speakers, with $U = I - 1$, and $\boldsymbol{\delta} = [\delta_1 \; \dots \; \delta_U]$ controls the amount of suppression of the interfering speakers. This optimization problem is the same as the optimization problem of the conventional LCMP beamformer [205], but with different RTF vectors and covariance matrix. Therefore the wLCMP convolutional beamformer can be obtained as

$$\bar{\mathbf{w}}_{\text{wLCMP}} = -\mathbf{G}\mathbf{q}_{\text{wLCMP}}, \tag{4.14}$$

where the dereverberation matrix $\mathbf{G}$ is obtained as in (4.11) and the beamforming vector $\mathbf{q}_{\text{wLCMP}}$ is obtained as in [205], i.e.,

$$\mathbf{q}_{\text{wLCMP}} = \mathbf{R}_d^{-1} \bar{\mathbf{C}} \left(\bar{\mathbf{C}}^H \mathbf{R}_d^{-1} \bar{\mathbf{C}} \right)^{-1} \mathbf{p}, \tag{4.15}$$

with $\bar{\mathbf{C}} = \begin{bmatrix} \bar{\mathbf{a}} & \bar{\mathbf{B}} \end{bmatrix}$ and $\mathbf{p} = \begin{bmatrix} 1 & \boldsymbol{\delta} \end{bmatrix}^T$. Setting δ_u to zero in (4.15) corresponds to a complete suppression of the u-th interfering speaker, while $\delta > 0$ leads to a controlled suppression.

The RETF vector of the target speaker $\bar{\mathbf{a}}$ in (4.15) is estimated using (4.12). The RETF vector of the u-th interfering speaker $\bar{\mathbf{b}}_u$ is estimated as

$$\bar{\mathbf{b}}_u = \mathbf{R}_{\sim_{u+v}} \text{MaxEig}\left(\mathbf{R}_{\sim_{u+v}}^{-1} \mathbf{R}_u \right) \tag{4.16}$$

where $\mathbf{R}_u = \frac{\sum_k \gamma_{u,k} \mathbf{d}_k \mathbf{d}_k^H}{\sum_k \gamma_{u,k}}$ is the covariance matrix of the u-th interfering speaker and $\mathbf{R}_{\sim_{u+v}} = \frac{\sum_k (1-\gamma_{u,k}) \mathbf{d}_k \mathbf{d}_k^H}{\sum_k (1-\gamma_{u,k})}$.

The output signal of the wLCMP convolutional beamformer is iteratively updated similarly as for the wMPDR convolutional beamformer.

4.3.3 *Relation with conventional MPDR and LCMP beamformers*

The conventional MPDR beamformer aims at minimizing the PSD of the output signal while preserving the reverberant target speech component in one of the microphone signals [10]. The MPDR beamformer is given by

$$\mathbf{w}_{\text{MPDR}} = \frac{\mathbf{R}_y^{-1}\mathbf{a}}{\mathbf{a}^H\mathbf{R}_y^{-1}\mathbf{a}}, \tag{4.17}$$

where $\mathbf{R}_y = \frac{1}{K}\sum_k \mathbf{y}_k\mathbf{y}_k^H$ and $\mathbf{a}$ denotes the reverberant RTF vector corresponding to the target speaker. The MPDR beamformer in (4.17) is similar to the convolutional wMPDR beamformer in (4.11) except that the covariance matrix $\mathbf{R}_y$ and the RTF vector $\mathbf{a}$ are estimated using the microphone signals $\mathbf{y}_k$ instead of the dereverberated microphone signals $\mathbf{d}_k$. In addition, the MPDR beamformer is obtained using a non-iterative optimization procedure compared to the wMPDR convolutional beamformer.

A similar relation exists between the conventional LMCP beamformer incorporating interfering speaker suppression constraints and the wLMCP convolutional beamformer in (4.15). The conventional LCMP beamformer is given by [205]

$$\mathbf{w}_{\text{LCMP}} = \mathbf{R}_y^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{R}_y^{-1}\mathbf{C}\right)^{-1}\mathbf{p}, \tag{4.18}$$

with $\mathbf{C} = [\mathbf{a} \quad \mathbf{B}]$ and $\mathbf{B} = [\mathbf{b}_1 \ \ldots \ \mathbf{b}_U]$ containing the reverberant RTF vectors of U interfering speakers.

The output signals of the MPDR and the LCMP beamformer are obtained as

$$z_k = \mathbf{w}_{\{\text{MPDR, LCMP}\}}^H\mathbf{y}_k. \tag{4.19}$$

These output signals are obviously computed without involving a dereverberation step compared to the output signals of wMPDR and wLCMP convolutional beamformers in (4.6).

4.4 Experimental setup

4.4.1 *Acoustic simulation setup*

In the experimental evaluation we consider two competing speakers, i.e., $I = 2$. Two German audio stories, uttered by two different male speakers, were used as the clean speech signals $s_1[n]$ and $s_2[n]$. Speech pauses that exceeded 0.5 s were shortened to 0.5 s, resulting in two highly overlapping (competing) audio stories. The hearing aid microphone signals $y_m[n]$ were generated at a sampling frequency of 16 kHz by convolving the clean speech signals with non-individualized measured binaural impulse responses (anechoic or reverberant) from [110], and adding diffuse babble noise, simulated according to [192]. The hearing aid setup in [110] consisted of two hearing aids, each equipped with three microphones ($M = 6$), mounted on

a dummy head. The left and the right competing speaker were simulated at $\theta_1 = -45^\circ$ and $\theta_2 = 45^\circ$. We consider three acoustic conditions, i.e., an anechoic-noisy condition with an average frequency-weighted segmental SNR (fwSSNR) of 2.9 dB, a reverberant condition (reverberation time $T_{60} \approx 0.5$ s) with an average fwSSNR of 3.5 dB, and a reverberant-noisy condition with an average fwSSNR of 0.5 dB. The average fwSSNR is computed by averaging the highest fwSSNR corresponding to speaker 1 and to speaker 2 among the microphone signals. The reference signals used to compute the fwSSNR are the anechoic speech signals $x_{i,m}^{an}[n]$ of the speakers at the first microphone of the hearing aid located at the same side of each speaker.

4.4.2 *Mask estimation*

The mask estimation neural network consisted of 3 BLSTM layers of 896 units. The network was trained on simulated noisy and reverberant mixtures obtained by mixing Librispeech [211] utterances convolved with room impulse responses generated with the image method for reverberation times between 0.2 s and 0.6 s, and adding babble noise at SNRs between 5 and 15 dB. The number of training mixtures was 50k. Note that there is a large mismatch between the training and the testing condition with respect to reverberation, background noise and head shadow effect, and also a large linguistic dissimilarity, as Librispeech consists of English read speech but the test data consists of German audio stories.

4.4.3 *Beamforming*

All considered beamformers were implemented using a weighted overlap-add (WOLA) framework with an STFT frame length FL = 512, an overlap of 75% between successive frames and a Hann window. For the wMPDR and wLCMP convolutional beamformers, the frame delay b was set to 4 and the length of the dereverberation filter was set to $L_w = 20$, 16 and 8 for frequency ranges $0-0.8$kHz, $0.8-1.5$kHz and $1.5-3$kHz, respectively. The variance of the target speech component was initialized as $\lambda_k = \|\mathbf{y}_k\|^2$. For the wLCMP convolutional beamformer and the LCMP beamformer, we set the interference suppression parameter to $\delta = 0.1$ to partially suppress the unattended speaker. The outputs signal of the wMPDR and wLCMP convolutional beamformers were obtained with 10 iterations.

To investigate the impact of mask estimation errors on the speech enhancement performance of the proposed system, we consider oracle ideal ratio masks (oMASK) and estimated ideal ratio masks (eMASK), obtained by the mask estimation neural network in (4.3).

We also compare our proposed system with a state-of-the-art cognitive-driven system proposed in [31], which uses either a conventional MVDR beamformer or a conventional LCMV beamformer to generate reference signals. Contrary to the MPDR and LCMP beamformers described in Section 4.3.3, these MVDR and LCMV beamformers use a diffuse noise covariance matrix instead of $\mathbf{R}_y$ and are steered using estimated anechoic RTF vectors (based on estimated DOAs of both speakers) instead of estimated reverberant RTF vectors. For the LCMV beamformer, the interference

suppression parameter was set to $\delta = 0.1$. Similarly as in [31], the DOAs of both speakers were estimated using a classification-based method [90] and the anechoic RTF vectors corresponding to the estimated DOAs were selected from a database of (measured) prototype RTF vectors [110].

4.4.4 *Speaker selection using AAD*

We used EEG responses recorded for 16 native German-speaking participants, where 8 participants were instructed to attend to the left speaker and 8 participants to the right speaker. See [31] for details about the EEG recording and the AAD training and decoding configuration.

For the AAD training and decoding steps (see Section 4.2.4), the EEG recordings were split into 30-second trials, resulting in 40 trials for the anechoic-noisy condition as well as for the reverberant-noisy condition, and 20 trials for the reverberant condition. Each participant's own data were used for training the spatio-temporal filter used for reconstructing the speech envelope $\hat{e}_a[l]$ from the EEG data.

4.4.5 *Performance measures*

We evaluate the cognitive-driven beamformers both in terms of AAD and speech enhancement performance. To evaluate the AAD performance, a trial is considered to be correctly decoded if the fwSSNR corresponding to the selected beamformer output signal $\hat{x}_a[n]$ (as the attended speech signal) is larger than the fwSSNR corresponding to the discarded beamformer output signal. To compute fwSSNR, the anechoic speech component $x_{a,m}^{an}[n]$ of the attended speaker in the first microphone signal of the hearing aid at the side of the attended speaker was used as the fwSSNR reference signal. The AAD performance is then computed by averaging the percentage of correctly decoded trials over all considered trials and all participants.

The speech enhancement performance of the complete proposed system is evaluated in terms of the fwSSNR improvement (ΔfwSSNR) using the same reference signals as used for AAD performance evaluation. The input fwSSNR is defined as the highest fwSSNR among the microphone signals. The output fwSSNR is defined as the fwSSNR of the selected beamformer output signals $\hat{x}_a[n]$.

To investigate the impact of the errors of speaker selection using AAD on the speech enhancement performance of the complete proposed system, we will consider oracle AAD (oAAD) where the attended speech signal $\hat{x}_a[n]$ is determined based on the highest ΔfwSSNR among the output signals of BEAM_1 and BEAM_2, and estimated AAD (eAAD) where $\hat{x}_a[n]$ is determined based on the highest Pearson correlation coefficients as described in Section4.2.4.

4.5 Experimental results

In this section, we evaluate the AAD performance and the speech enhancement performance of the proposed cognitive-driven convolutional beamforming system.

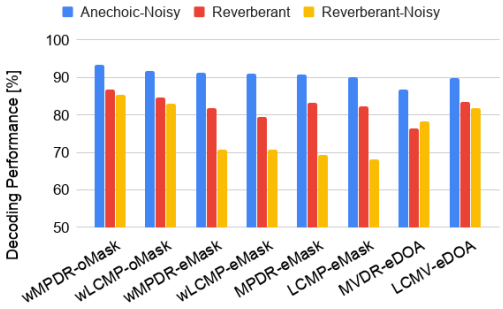


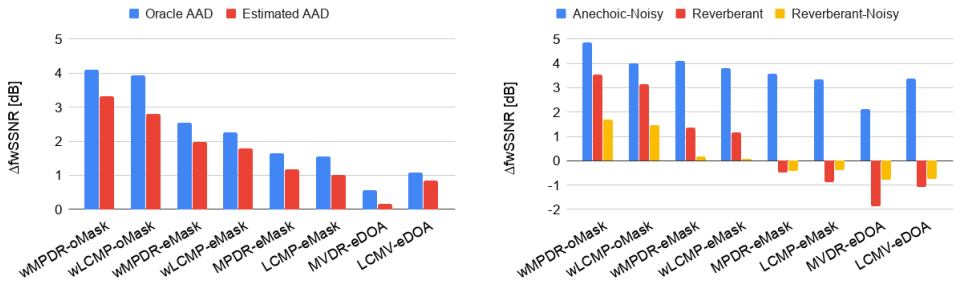
Fig. 4.2: Average auditory attention decoding performance for the anechoic-noisy, reverberant and reverberant-noisy conditions for the different considered beamformers. The upper boundary of the confidence interval corresponding to chance level for the anechoic-noisy, reverberant and reverberant-noisy conditions are 61.39%, 66.19%, 61.39%, respectively, computed based on a binomial test at the 5% significance level.

In Section 4.5.1 we investigate the impact of mask estimation errors on the AAD performance. In Section 4.5.2, we investigate the impact of AAD errors on the speech enhancement performance.

4.5.1 *Auditory attention decoding performance*

Figure 4.2 depicts the average AAD performance for the anechoic-noisy, the reverberant and the reverberant-noisy condition, when using the output signals of the wMPDR or wLCMP convolutional beamformer, the MPDR or LCMP beamformer and the MVDR or LCMV beamformer as reference signals for decoding. We observe that all considered beamformers yield a AAD performance that is significantly larger than chance levels. For all considered acoustic conditions the wMPDR convolutional beamformer and the wLCMP convolutional beamformer using the oracle masks (wMPDR-oMASK and wLCMP-oMASK) yield the highest AAD performance, showing the potential of using convolutional beamformers for AAD.

When using estimated masks instead of oracle masks for the convolutional beamformers (wMPDR-eMASK and wLCMP-eMASK) the AAD performance decreases, especially in the reverberant-noisy condition. In the reverberant-noisy condition, the MVDR and LCMV beamformers using anechoic RTF vectors based on estimated DOAs (MVDR-eDOA and LCMV-eDOA) yield a larger average AAD performance than the beamformers using reverberant RTF vectors based on the estimated masks. This suggests that in order to improve the AAD performance, a better estimation of RTF vectors is required, e.g., based on prototype RTF vectors or neural networks that are more robust to background noise and reverberation.



(a) (b)

Fig. 4.3: fwSSNR improvement (a) averaged over all considered acoustic conditions when using oracle AAD and estimated AAD (b) for the anechoic-noisy, reverberant and reverberant-noisy conditions when using estimated AAD. The input fwSSNR averaged over all considered acoustic conditions is 2.06dB and the input fwSSNRs for the anechoic-noisy, reverberant and reverberant-noisy conditions are 2.9dB, 3.5dB, 0.5dB, respectively.

4.5.2 *Speech enhancement performance*

Figure 4.3a depicts the fwSSNR improvement of the complete proposed system averaged over all considered acoustic conditions, either using oracle AAD or estimated AAD. It can be observed that the convolutional beamformers outperform all other considered beamformers for both oracle and estimated AAD. When using estimated AAD instead of oracle AAD, for all considered beamformers the fwSSNR improvement decreases by 0.2–1.1 dB, showing the sensitivity to AAD errors. Nevertheless, the fwSSNR improvement of the convolutional beamformers is about 1.6–1.8 dB larger than the state-of-the-art MVDR and LCMV beamformers using estimated DOAs.

Figure 4.3b depicts the fwSSNR improvement of the complete proposed system for the anechoic-noisy, reverberant and reverberant-noisy conditions when using estimated AAD. It can be observed that all beamformers yield a significant fwSSNR improvement for the anechoic-noisy condition. However, for the reverberant condition the systems using conventional beamformers (MPDR-eMask, LCMP-eMask, MVDR-eDOA, LCMV-eDOA) tend to degrade the fwSSNR, whereas only the proposed system using convolutional beamformers (wMPDR-eMask, wLCMP-eMask) provides a fwSSNR improvement, showing the influence of dereverberation. It should be noted that the considered reverberant-noisy condition with an interfering speaker is an extremely adverse condition with babble noise at a signal-to-interference-plus-noise ratio (SINR) of 0.3 dB and a reverberation time of 0.5 s, which makes it very challenging for speech enhancement.

4.5.3 *Discussion*

The experimental results show that for the considered acoustic setup the AAD performance and the fwSSNR improvement of the proposed cognitive-driven speech enhancement system using convolutional beamformers are sensitive to mask estimation errors, particularly for the reverberant and reverberant-noisy conditions. The mask estimation errors can be mainly attributed to the linguistic dissimilarity of training and testing conditions of the neural-network-based mask estimation algorithm and also the intrinsic difficulty of separating out two competing speakers with the same gender in the reverberant-noisy condition.

The results show that the wMPDR convolutional beamformer yields a larger fwSSNR improvement than the wLCMP convolutional beamformer. Although the wMPDR convolutional beamformer can strongly suppress the interfering speaker, it may deprive the listener from the ability to switch attention between the speakers. In contrast, the wLCMP convolutional beamformer is able to both control the interfering speaker suppression as well as yield a considerable fwSSNR improvement.

Lastly, the results show that the convolutional beamformers (wLCMP-eMASK and wMPDR-eMASK) yield the highest fwSSNR improvement for all considered acoustic conditions, whereas the conventional LCMV beamformer (LCMV-eDOA) yields the highest AAD performance in the reverberant and reverberant-noisy conditions. Future work could therefore investigate the potential of combining the convolutional and the conventional beamformers to improve both the decoding and the speech enhancement performance.

4.6 Conclusion

In this paper, we proposed a cognitive-driven speech enhancement system which combines neural-network-based mask estimation, convolutional beamformers and AAD. We considered the wMPDR convolutional beamformer, which jointly enhances the attended speaker and suppresses the unattended speaker, reverberation and background noise. In addition, we proposed a wLCMP convolutional beamformer which enables to control the amount of suppression for the unattended speaker. The experimental results showed that the proposed system using convolutional beamformers is able to considerably improve the fwSSNR both for noisy and reverberant conditions compared to state-of-the-art cognitive-driven speech enhancement systems.

5

CLOSED-LOOP COGNITIVE-DRIVEN GAIN CONTROL OF COMPETING SOUNDS USING AUDITORY ATTENTION DECODING

Recent advances in EEG have shown that it is possible to identify the target speaker which a listener is attending to using single-trial EEG-based auditory attention decoding (AAD) methods. However, the performance of most AAD methods has been investigated for an open-loop scenario, where AAD is performed in an off-line fashion without presenting on-line feedback to the listener. Aiming at developing a closed-loop AAD system that allows to enhance a target speaker, suppress an interfering speaker and switch attention between both speakers, in this paper we propose a cognitive-driven adaptive gain controller (AGC) based on real-time AAD. Using the EEG responses of the listener and the speech signals of both speakers, the real-time AAD generates probabilistic attention measures, based on which the attended and the unattended speaker are identified. The AGC then amplifies the identified attended speaker and attenuates the identified unattended speaker, which are presented to the listener via loudspeakers. We investigate the performance of the proposed system in terms of the decoding performance and the signal-to-interference ratio (SIR) improvement. The experimental results show that closed-loop AAD in an on-line fashion is feasible and enables the listener to interact with the AGC. Furthermore, the results show that although there is a significant delay to detect attention switches, the proposed system is able to improve the SIR between the attended and the unattended speaker. In addition, no significant difference in decoding performance is observed between closed-loop AAD and open-loop AAD. The subjective evaluation results show that the proposed closed-loop cognitive-driven system demands a similar level of cognitive effort to follow the attended speaker, to ignore the unattended speaker and to switch attention between both speakers compared to using open-loop AAD.

5.1 Introduction

Hearing aids aim at restoring the normal hearing abilities by several processing steps including speech enhancement. The main objective of speech enhancement is to improve the intelligibility of the recorded microphone signals, which are often

corrupted by various noise sources [10, 13]. In a scenario with multiple competing speakers the performance of many speech enhancement algorithms depends on correctly identifying the target speaker, i.e., the speaker which the listener is attending to.

Recent advances in electroencephalography (EEG) have shown that it is possible to identify the target speaker from single-trial EEG recordings [16, 20–26, 128, 130, 163, 172]. Several single-trial EEG-based auditory attention decoding (AAD) methods to identify the attended speaker have been proposed based on, e.g., a least-squares cost function [16, 20, 22, 25, 26, 163], canonical correlation analysis [21, 128], a state-space model [24, 172] and neural networks [23, 130]. The least-squares-based AAD method used in [16, 20, 22, 25, 26, 163] aims at reconstructing the attended speech envelope from the EEG responses of the listener using a trained spatio-temporal envelope estimator. To identify the attended speaker, the reconstructed speech envelope is compared with the speech envelopes of the competing speakers using correlation coefficients. Since these correlation coefficients are typically highly fluctuating, a large correlation window on the order of about 30 seconds is typically required to obtain a reliable decoding performance, which causes a large processing delay.

The possibility of decoding auditory attention from EEG recordings has led to an increasing research interest in the topic of incorporating AAD in a brain-computer interface for real-world applications, e.g., to cognitively drive speech enhancement algorithms [27, 30, 31, 198, 212]. Cognitive-driven speech enhancement algorithms potentially provide the listener with the ability to selectively attend to a specific speaker. It should however be noted that the performance of most aforementioned AAD methods and cognitive-driven speech enhancement algorithms has been investigated for an open-loop scenario, where AAD is performed in an off-line fashion without presenting on-line feedback to the listener. In addition, scenarios with no attention switch between speakers have typically been investigated, which is unrealistic in practice. To investigate the performance of AAD for real-world applications, closing the loop by presenting feedback according to the AAD results in an on-line fashion is of crucial importance. Feedback presentation may influence the subsequent intent of the listener and the brain signals that encode that intent. In [163] the feasibility of a closed-loop system based on least-squares-based AAD has been shown by presenting the AAD results as visual feedback, i.e., using different colors or a sphere with different radii. However, the feasibility of closed-loop AAD enabling the listener to interact with speech enhancement in an on-line fashion and allowing the listener to switch attention between speakers remains to be investigated.

Aiming at developing a closed-loop AAD system that allows to enhance a target speaker, suppress an interfering speaker and switch attention between both speakers, we propose a cognitive-driven adaptive gain controller (AGC), which is based on real-time AAD (RAAD). The RAAD generates correlation coefficients for both speakers using the EEG responses of the listener and the speech signals of both speakers. These correlation coefficients are then translated into more reliable probabilistic attention measures, based on which the attended and the unattended speaker are identified. The AGC as an ideal speech enhancement algorithm then amplifies the identified attended speaker and attenuates the identified unattended speaker. Fi-

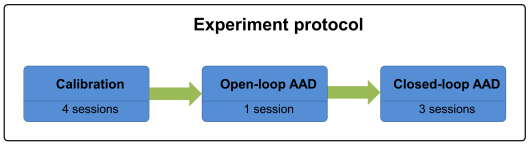


Fig. 5.1: Experiment protocol used to calibrate and evaluate the cognitive-driven gain controller system. The experiment protocol consists of a calibration phase with four sessions, an open-loop AAD phase with one session and a closed-loop AAD phase with three sessions.

nally, the loop of cognitive-driven gain control is closed by presenting the amplified attended speaker and the attenuated unattended speaker via loudspeakers.

To generate correlation coefficients of speakers, we adopt the least-squares-based AAD method from [16]. The correlation coefficients are generated either using a small correlation window of length 0.25 seconds or using a large correlation window of length 15 seconds. To identify the attended and the unattended speaker from the fluctuating correlation coefficients, we propose an AAD algorithm using a generalized linear model (GLM) and consider an AAD algorithm using a state-space model (SSM), as proposed in [24, 172]. To amplify the attended speaker and attenuate the unattended speaker, we propose an AGC which adjusts the gains of both speakers based on the probabilistic attention measures, as proposed in [213]. For an acoustic scenario comprising two competing speakers we investigate the speech enhancement performance of the proposed closed-loop cognitive-driven gain controller system based on objective and subjective evaluations. In addition, we provide a detailed analysis and experimental comparison between an open-loop AAD system and the closed-loop AAD system using either the GLM or the SSM.

5.2 Methods

5.2.1 *Experiment protocol*

In this section, we present the experiment protocol used to calibrate and evaluate the cognitive-driven gain controller system. The experiment protocol consists of a calibration phase, an open-loop AAD phase and a closed-loop AAD phase (see Fig. 5.1).

5.2.1.1 *Calibration phase*

In the calibration phase the cognitive-driven gain controller system was calibrated using the EEG responses for a scenario with two competing speakers (see Fig. 5.2). Participants were cued by an arrow on a screen to listen attentively to one of the speakers while recording the ongoing EEG responses. Participants were also instructed to minimize eye movement and blinking. The EEG recordings were recorded during four sessions, lasting 30 minutes in total. The first and the second session

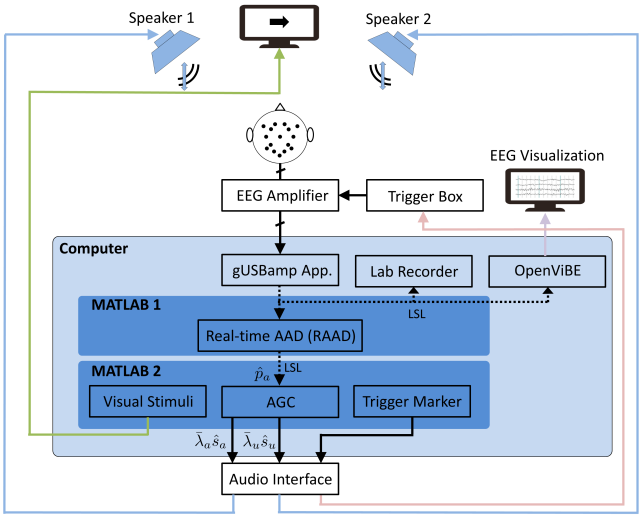


Fig. 5.2: Overview of the proposed cognitive-driven gain controller system for a scenario with two competing speakers. The EEG responses were acquired using the EEG amplifier. The acquired EEG responses and EEG trigger markers were streamed using the gUSBamp application. Using the gUSBamp application and the LSL software package, the streamed EEG responses were forwarded to the real-time AAD (RAAD) for on-line decoding, to the Lab Recorder application for recording, and to the OpenViBE software for on-line EEG visualization. The RAAD was implemented and run using MATLAB (MATLAB 1). The RAAD identified the attended and the unattended speaker and generated their corresponding probabilistic attention measure. The generated probabilistic attention measure of the attended speaker ($\hat{p}_a$) was forwarded to the AGC using the LSL software package. Based on the probabilistic attention measure, the AGC amplified the attended speaker ($\bar{\lambda}_a \hat{s}_a$) and attenuated the unattended speaker ($\bar{\lambda}_u \hat{s}_u$) as acoustic stimuli. The AGC (together with trigger marker and visual stimuli) were implemented and run using MATLAB (MATLAB 2). The AAD loop is then closed by presenting the acoustic stimuli using the audio interface and two loudspeakers.

each lasted 10 minutes, while the third and the fourth session each lasted 5 minutes. For the first and the third session the participants were cued to attend to the left speaker, whereas for the second and the fourth session the participants were cued to attend to the right speaker. Following each session, the participants were asked to fill out a questionnaire consisting of multiple-choice questions about the stories uttered by the speakers. There was one question per minute of each story. The questionnaire was used to check whether the participants attended to the cued speaker. After the fourth session, there was a short break. During this break, the recorded EEG responses were used to calibrate the cognitive-driven gain controller system (see Section 5.2.5), individualized per participant.

5.2.1.2 *Open-loop AAD phase*

In the open-loop AAD phase, the calibrated AAD algorithms were employed to identify the attended and the unattended speaker without presenting feedback to the participants. Similarly to the calibration phase, the scenario comprised two competing speakers. The open-loop AAD phase consisted of one session lasting 10 minutes. During this session, participants were cued by an arrow on a screen every minute to switch attention between the competing speakers while recording the ongoing EEG responses. Afterwards, participants were asked to rate how effortful it was to follow the attended speaker, to ignore the unattended speaker and to switch attention to the cued speaker on a scale from 0 to 10, with 0 being least effortful and 10 being most effortful. In addition, participants were asked to rate how well they understood the attended story on a scale from 0 to 10, with 0 being nothing understood and 10 being everything understood. While participants were rating, the decoding performance of several AAD algorithms (see Section 5.2.5.1) was evaluated using the recorded EEG responses. The following settings for the correlation window used in the AAD algorithms were considered:

- LW–GLM: AAD algorithm using generalized linear model with a large correlation window of 15 seconds[1].
- LW–SSM: AAD algorithm using state-space model with a large correlation window of 15 seconds.
- SW–SSM: AAD algorithm using state-space model with a small correlation window of 0.25 seconds. Using a small correlation window was motivated by the results in [24], where it was shown that the state-space model is able to translate highly fluctuating coefficients of the spatio-temporal envelope estimators into reliable probabilistic attention measures (see Section 5.2.5.1-D).

5.2.1.3 *Closed-loop AAD phase*

In the closed-loop AAD phase, the calibrated cognitive-driven gain controller system was employed to identify the attended and the unattended speaker and to close the loop by presenting the amplified attended speaker and the attenuated unattended speaker in an on-line fashion via loudspeakers. The closed-loop AAD phase consisted of three sessions, each lasting 10 minutes. During each session, participants were cued by an arrow on a screen every minute to switch attention between the presented competing speakers while recording the ongoing EEG responses. To identify the attended and the unattended speaker, in each session a different AAD algorithm was used, i.e., LW–GLM, LW–SSM and SW–SSM. These AAD algorithms were randomly assigned to the sessions. For analysing the experimental results, 24% of the results needed to be excluded, 5% due to a technical hardware problem

1 Note that in this paper an AAD algorithm using GLM with a small correlation window was not considered, since initial experiments showed highly fluctuating correlation coefficients with unreliable probabilistic attention measures.

when saving the results and 19% due to poor attentional performance reported by participants themselves after a few sessions[2].

5.2.2 *Participants*

Ten native German-speaking participants took part in this study. All participants were self-reported normal hearing and reported no past or present neurological or psychiatric conditions.

5.2.3 *Stimuli*

Two German audio stories, uttered by two different male speakers, were used as the speech signals of the competing speakers. One story was from a German audio book website [189] and the other story was from a selection of audio books [188]. Before performing the experiment, participants reported no knowledge of the audio stories. Speech pauses from the audio stories that exceeded 0.5 s were shortened to 0.5 s. The audio stories were normalized to the same root-mean-square (RMS) value at a comfortable level which was individualized by each participant. The audio stories with no repetition were considered as the acoustic stimuli for the calibration phase, the open-loop AAD phase and the closed-loop AAD phase. The acoustic stimuli were presented at a sampling frequency of 44100 Hz using MATLAB (MATLAB 2 in Fig. 5.2), a Fireface UC audio interface system (provided by RME Audio, Germany) and two loudspeakers placed at the left side (with an azimuth of $-45°$) and the right side (with an azimuth of $45°$) of the participants. The visual stimuli consisting of an arrow for cueing were presented using a monitor in front of the participants. In addition, the EEG trigger markers synchronized with the acoustic and visual stimuli were generated using the Fireface UC audio interface system and a g.TRIGbox (provided by g.tec, Austria). The presentation of the acoustic and visual stimuli and the trigger marker generation were performed using the same computer employed for the cognitive-driven gain controller system (see Fig. 5.2).

5.2.4 *Data acquisition*

Aiming at using a small number of electrodes for AAD, EEG responses were acquired using $C = 16$ electrodes. The electrodes were placed on the scalp area at F1, F2, FC3, FC4, FT7, FT8, Cz, C5, C6, P5, P4, P7, P8, Oz, PO3 and PO4. This electrode placement was inspired by the results in [20, 165], where it was shown that an electrode configuration covering the temporal, central, frontal and parental scalp areas yields a reliable decoding performance. The EEG responses were referenced to the P9 electrode. The EEG responses were acquired using active (g.LADYbird) electrodes and

2 Some participants reported that they either lost their concentration to attend to the cued speaker or they engaged with the story uttered by the non-cued speaker.

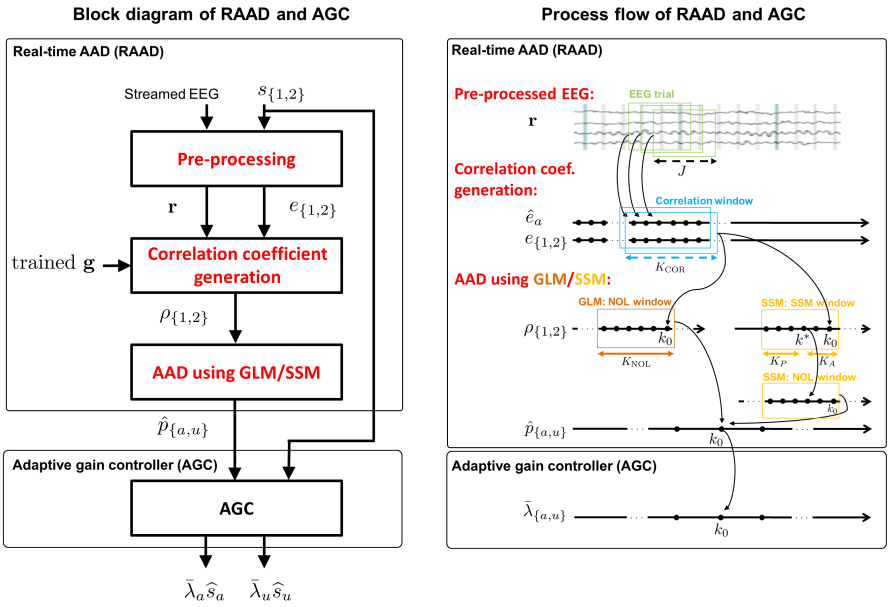


Fig. 5.3: Block diagram and process flow of the real-time AAD (RAAD) and the adaptive gain controller (AGC).

a g.USBamp bio-signal amplifier (provided by g.tec, Austria). The acquired EEG responses and EEG trigger markers were streamed at a sampling frequency of 500 Hz using the gUSBamp application from the Lab Streaming Layer (LSL) software package (provided by Swartz Center for Computational Neuroscience, UCSD). Using the gUSBamp application, the streamed EEG responses were also forwarded to RAAD for on-line decoding, to the Lab Recorder application (provided by Swartz Center for Computational Neuroscience and Kothe) for recording, and to the OpenViBE software for on-line EEG visualization. The gUSBamp application, the OpenViBE software and the Lab Recorder application were run on the same computer employed for the cognitive-driven gain controller system (see Fig. 5.2).

5.2.5 *Cognitive-driven gain controller system*

In this section, we present the proposed cognitive-driven gain controller system consisting of RAAD and AGC (see Fig. 5.3). Section 5.2.5.1 describes the RAAD, which generates probabilistic attention measures based on which the attended and the unattended speaker are identified. Section 5.2.5.2 describes the AGC, which amplifies the identified attended speaker and attenuates the identified unattended speaker based on the probabilistic attention measures.

5.2.5.1 *Real-time auditory attention decoding (RAAD)*

The RAAD consists of three blocks (see Fig. 5.3), i.e., pre-processing of the EEG responses, correlation coefficient generation and AAD using either GLM or SSM.

A. *Pre-processing:* The streamed EEG responses from the gUSBamp application were re-referenced to a common average reference, band-pass filtered between 0.5 Hz and 9 Hz using a fourth-order Butterworth band-pass filter, and subsequently downsampled to 64 Hz in an on-line fashion. Contrary to the on-line EEG pre-processing, the speech pre-processing was performed in an off-line fashion, i.e., the speech signal $s_{1,t}$ of speaker 1 and the speech signal $s_{2,t}$ of speaker 2, with t the discrete time index for $t = 1 \dots T$, were assumed to be available. The envelopes of both speech signals $e_{1,k}$ and $e_{2,k}$, with k the sub-sampled time index for $k = 1 \dots K$, were obtained using a Hilbert transform, followed by low-pass filtering at 9 Hz and downsampling to 64 Hz. The pre-processed EEG responses and the speech envelopes were then provided in an on-line fashion to the correlation coefficient generation block.

B. *Correlation coefficient generation:* To generate the correlation coefficients of speaker 1 and speaker 2, we adopted the least-squares-based AAD method from [16], which estimates the attended speech envelope from the EEG recordings using a spatio-temporal envelope estimator trained during the calibration phase.

   1) *Training step (calibration phase):* In the training step, the attended speaker is assumed to be known. The attended speech envelope is then estimated from the pre-processed EEG responses $r_{c,k}$, with c the electrode index for $1 \dots C$, using a spatio-temporal envelope estimator $\mathbf{g}$ [16], i.e.,

$$\hat{e}_{a,k} = \mathbf{g}^T \mathbf{r}_k, \tag{5.1}$$

   with

$$\mathbf{g} = \left[\mathbf{g}_1^T \, \mathbf{g}_2^T \, \dots \, \mathbf{g}_C^T\right]^T, \tag{5.2}$$

$$\mathbf{g}_c = \left[g_{c,0} \;\; g_{c,1} \dots \; g_{c,J-1}\right]^T, \tag{5.3}$$

$$\mathbf{r}_k = \left[\mathbf{r}_{1,k}^T \, \mathbf{r}_{2,k}^T \; \dots \; \mathbf{r}_{C,k}^T\right]^T, \tag{5.4}$$

$$\mathbf{r}_{c,k} = \left[r_{c,k} \, r_{c,k+1} \; \dots \; r_{c,k+J-1}\right]^T, \tag{5.5}$$

   where J denotes the number of envelope estimator coefficients per electrode. The trained envelope estimator $\mathbf{g}$ is obtained by minimizing the least-squares error between the (known) attended speech envelope $e_{a,k}$ and the reconstructed envelope $\hat{e}_{a,k}$, regularized with the squared-norm of the derivative of the envelope estimator coefficients to avoid overfitting [16, 26, 129], i.e.,

$$\min_{\mathbf{g}} \frac{1}{K} \sum_{k=1}^{K} (e_{a,k} - \underbrace{\mathbf{g}^T \mathbf{r}_k}_{\hat{e}_{a,k}})^2 + \beta \mathbf{g}^T \mathbf{\Lambda} \mathbf{g}, \tag{5.6}$$

with $\mathbf{\Lambda}$ denoting the derivative matrix [26] and β denoting a regularization parameter. The solution of (5.6) is equal to

$$\mathbf{g} = (\mathbf{Q} + \beta\mathbf{\Lambda})^{-1}\,\mathbf{q}, \tag{5.7}$$

with the correlation matrix $\mathbf{Q}$ and the cross-correlation vector $\mathbf{q}$ given by

$$\mathbf{Q} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{r}_k\mathbf{r}_k^T, \quad \mathbf{q} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{r}_k e_{a,k}. \tag{5.8}$$

2) *Correlation coefficient generation step (open-loop and closed-loop AAD phase):* To generate the correlation coefficients of speaker 1 and speaker 2, we compute the Pearson correlation coefficients between the estimated attended envelope $\hat{e}_{a,k}$ in (5.1) and the speech envelopes $e_{1,k}$ and $e_{2,k}$, i.e.,

$$\rho_{1,k} = \rho\left(\mathbf{e}_{1,k},\ \hat{\mathbf{e}}_{a,k}\right), \quad \rho_{2,k} = \rho\left(\mathbf{e}_{2,k},\ \hat{\mathbf{e}}_{a,k}\right), \tag{5.9}$$

where $\hat{\mathbf{e}}_{a,k}$ denotes the stacked vector of estimated attended envelopes corresponding to a correlation window of length K_{COR}, i.e.,

$$\hat{\mathbf{e}}_{a,k} = \left[\hat{e}_{a,k-K_{\mathrm{COR}}+1}\ \hat{e}_{a,k-K_{\mathrm{COR}}+2}\ \ldots\ \hat{e}_{a,k}\right]^T, \tag{5.10}$$

and $\mathbf{e}_1[k]$ and $\mathbf{e}_2[k]$ are defined similarly as in (5.10).

In the training step, the pre-processed EEG responses obtained from the calibration phase were segmented into trials of length 15 seconds, shifted by 1 sample (corresponding to $\frac{1}{64}$ seconds). In addition, the parameters J and β of the envelope estimator in (5.3) and (5.7) were determined for each participant using a leave-one-trial-out cross-validation approach, similarly as in [16, 26]. Using these parameters, an envelope estimator $\mathbf{g}$ in (5.7) for each participant was then computed using all trials from the calibration phase.

In the correlation coefficient generation step, the EEG responses were segmented in the same way as in the training step. The correlation coefficients were computed either using a large correlation window of length $K_{\mathrm{COR}} = 960$ samples (corresponding to 15 seconds) with an overlap of 959 samples or using a small correlation window of length $K_{\mathrm{COR}} = 16$ samples (corresponding to 0.25 seconds) with no overlap. In [18, 22, 24, 172] it has been shown that the performance of AAD algorithms is affected by fluctuations of the correlation coefficients $\rho_{1,k}$ and $\rho_{2,k}$ in (5.9). In this paper we propose to use two methods (GLM and SSM) to translate the fluctuating correlation coefficients into more reliable probabilistic attention measures.

C. *Auditory attention decoding using generalized linear model:* The AAD algorithm using the GLM consists of a training and a decoding step. The training step takes place during the calibration phase, whereas the decoding step takes place during the open-loop and the closed-loop AAD phase.

1) *Training step:* To estimate the probability of attending to speaker 1 or speaker 2, the correlation coefficients of speaker 1 and speaker 2 in (5.9)

are first segmented into non-overlapping (NOL) windows of length K_{NOL}, i.e.,

$$\boldsymbol{\rho}_{1,i} = \left[\rho_{1,(i-1)K_{\mathrm{NOL}}+1}\ \rho_{1,(i-1)K_{\mathrm{NOL}}+2}\ \cdots\ \rho_{1,iK_{\mathrm{NOL}}}\right]^T, \tag{5.11}$$

$$\boldsymbol{\rho}_{2,i} = \left[\rho_{2,(i-1)K_{\mathrm{NOL}}+1}\ \rho_{2,(i-1)K_{\mathrm{NOL}}+2}\ \cdots\ \rho_{2,iK_{\mathrm{NOL}}}\right]^T, \tag{5.12}$$

The mean differential correlation coefficient of speaker 1 and speaker 2 in window i is computed as

$$\bar{\Delta}\rho_i = \frac{1}{K_{\mathrm{NOL}}} \sum_{n=1}^{K_{\mathrm{NOL}}} (\rho_{1,(i-1)K_{\mathrm{NOL}}+n} - \rho_{2,(i-1)K_{\mathrm{NOL}}+n}). \tag{5.13}$$

We model the attention state $\bar{d}_i$ in window i as a binary random variable [197], i.e.,

$$\begin{cases} \bar{d}_i = 1, & \text{attending to speaker 1 in window } i \\ \bar{d}_i = 2, & \text{attending to speaker 2 in window } i \end{cases}, \tag{5.14}$$

which is assumed to follow a Bernoulli distribution with probability $\bar{p}_i$, i.e.,

$$P\left(\bar{d}_i\right) = \bar{p}_i{}^{2-\bar{d}_i} \left(1-\bar{p}_i\right)^{\bar{d}_i-1} = \begin{cases} \bar{p}_i, & \text{if } \bar{d}_i{=}1 \\ 1-\bar{p}_i, & \text{if } \bar{d}_i{=}2 \end{cases}. \tag{5.15}$$

Using a GLM, the probability of attending to speaker 1 is given by [214]

$$\bar{p}_i = P\left(\bar{d}_i = 1\right) = 1 - P\left(\bar{d}_i = 2\right) = \frac{1}{1+e^{-\bar{z}_i}}, \tag{5.16}$$

with the linear predictor $\bar{z}_i$, i.e.,

$$\bar{z}_i = \mathbf{x}_i^T \boldsymbol{\alpha}, \tag{5.17}$$

$$\mathbf{x}_i = [1\ \ \bar{\Delta}\rho_i]^T, \tag{5.18}$$

$$\boldsymbol{\alpha} = [\alpha_0\ \ \alpha_1]^T, \tag{5.19}$$

where α_0 and α_1 denote the GLM parameters. Obviously, the probability of attending to speaker 1 monotonically increases from 0 to 1 for $\bar{z}_i \in (-\infty, \infty)$.

To obtain the GLM parameters α_0 and α_1, the probability mass function in (5.15) can be written as an exponential distribution using the canonical link function $\theta_i = \mathrm{logit}\left(\bar{p}_i\right) = \bar{z}_i$, with $\mathrm{logit}\left(\bar{p}_i\right) = \log\left(\frac{\bar{p}_i}{1-\bar{p}_i}\right)$, i.e.,

$$P\left(\bar{d}_i\right) = \exp(\bar{d}_i\theta_i - b\left(\theta_i\right)), \tag{5.20}$$

Algorithm 1 : GLM Training

input: $\mathbf{x}_i$ and $\bar{d}_i$ for $i = 1 \dots I$

1: initialization: $\bar{p}_i^{(0)} = \bar{d}_i$, $\bar{z}_i^{(0)} = \text{logit}(\bar{p}_i^{(0)})$, calculate $\hat{\boldsymbol{\alpha}}^{(1)}$ using (5.23)
2: **for** $r = 1 \dots R$ **do**
3: calculate $\bar{p}_i^{(r)}$ and $\bar{z}_i^{(r)}$ using (5.26) and (5.27), respectively
4: calculate $\mathbf{y}^{(r)}$ and $\mathbf{W}^{(r)}$ using (5.28), (5.29) and (5.25), respectively
5: update the GLM parameters $\hat{\boldsymbol{\alpha}}^{(r+1)}$ using (5.23)
6: **end for**

output: $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}^{(R+1)}$

with

$$b\left(\theta_i\right) = \log\left(1 + \exp\left(\theta_i\right)\right). \tag{5.21}$$

The GLM parameters in (5.19) are then obtained by maximizing the log-likelihood function, i.e.,

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \ell\left(\mathbf{x}_i, i = 1 : I | \boldsymbol{\alpha}\right) = \sum_{i=1}^{I} \bar{d}_i \theta_i - b\left(\theta_i\right), \tag{5.22}$$

which is a maximum likelihood (ML) estimate and can be computed by using an iteratively re-weighted least-squares algorithm and Newton-Raphson method as [215, 216]

$$\hat{\boldsymbol{\alpha}}^{(r+1)} = (\mathbf{X}^T \mathbf{W}^{(r)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(r)} \mathbf{y}^{(r)}, \tag{5.23}$$

with r the iteration index and

$$\mathbf{X} = [\mathbf{x}_1^T \, \mathbf{x}_2^T \dots \, \mathbf{x}_I^T]^T, \;\; \mathbf{X} \in \mathbb{R}^{I \times 2}, \tag{5.24}$$

$$\mathbf{W}^{(r)} = \text{diag}\left\{ \frac{1}{\bar{p}_i^{(r)}\left(1 - \bar{p}_i^{(r)}\right)} \right\}, \;\; \mathbf{W}^{(r)} \in \mathbb{R}^{I \times I}, \tag{5.25}$$

$$\bar{p}_i^{(r)} = \text{logit}^{-1}(\bar{z}_i^{(r)}), \tag{5.26}$$

$$\bar{z}_i^{(r)} = \mathbf{x}_i^T \hat{\boldsymbol{\alpha}}^{(r)}, \tag{5.27}$$

$$\mathbf{y}^{(r)} = [y_1^{(r)} \, y_2^{(r)} \dots \, y_I^{(r)}]^T, \;\; \mathbf{y}^{(r)} \in \mathbb{R}^{I \times 1}, \tag{5.28}$$

$$y_i^{(r)} = \bar{z}_i^{(r)} + (\bar{d}_i - \bar{p}_i^{(r)}) \text{logit}'(\bar{p}_i^{(r)}), \tag{5.29}$$

where $(\cdot)'$ denotes the derivative operator. Algorithm 1 summarizes the GLM parameter estimation in the training step.

2) *Decoding step:* To decode which speaker a participant is attending to in window i, the mean differential correlation coefficient $\bar{\Delta}\rho_i$ is computed using (5.13), based on which the linear predictor $\bar{z}_i$ is computed using the (trained) GLM parameters $\hat{\boldsymbol{\alpha}}$ in (5.17). The probability of attending

to speaker 1 $P\left(\bar{d}_i = 1\right)$ and the probability of attending to speaker 2 $P\left(\bar{d}_i = 2\right)$ are then obtained using (5.16). Based on these probabilities, it is decided that the participant attended to speaker 1 if $P\left(\bar{d}_i = 1\right) > P\left(\bar{d}_i = 2\right)$ or attended to speaker 2 otherwise. The probabilistic attention measure of the attended speaker $\hat{p}_{a,i}$ in window i is hence determined as

$$\begin{cases} \hat{p}_{a,i} = P\left(\bar{d}_i = 1\right), & \text{if } P\left(\bar{d}_i = 1\right) > P\left(\bar{d}_i = 2\right) \\ \hat{p}_{a,i} = P\left(\bar{d}_i = 2\right), & \text{otherwise.} \end{cases} \tag{5.30}$$

Obviously, the probabilistic attention measure of the attended speaker $\hat{p}_{a,i}$ can vary between 0.5 and 1. The probabilistic attention measure of the unattended speaker $\hat{p}_{u,i}$ is determined as $\hat{p}_{u,i} = 1 - \hat{p}_{a,i}$. The process flow of AAD using the GLM is depicted in Fig. 5.3.

The AAD algorithm using the GLM was implemented and run using MATLAB (MATLAB 1 of RAAD in Fig. 5.2). For the training step, Algorithm 1 with $R = 30$ iterations and the correlation coefficients obtained from the calibration phase were used. Both for the training and the decoding steps, the correlation coefficients were computed using the large correlation window (i.e., $K_{\text{COR}} = 960$ samples) and the mean differential correlation coefficients in (5.13) were generated using a window of length $K_{\text{NOL}} = 16$ samples (corresponding to 0.25 seconds). During the decoding step, the probabilistic attention measures $\hat{p}_{a,i}$ and $\hat{p}_{u,i}$ were forwarded to the AGC using the LSL software package (see Fig. 5.2). Each participant's own data were used for training and decoding. To evaluate the performance of the proposed LW–GLM algorithm, the decoding performance for each participant was computed as the percentage of correctly decoded NOL windows.

D. *Auditory attention decoding using state-space model:* As an alternative to the GLM, it has been proposed in [24] to use a SSM to translate the absolute values of the coefficients of the spatio-temporal envelope estimator into probabilistic attention measures. Contrary to [24], in this paper we propose to use the absolute values of the correlation coefficients

$$\phi_{1,k} = |\rho_{1,k}|, \tag{5.31}$$

$$\phi_{2,k} = |\rho_{2,k}|, \tag{5.32}$$

instead of the coefficients of the spatio-temporal envelope estimator, which need to be obtained for both the attended and the unattended speaker.

Similarly to (5.14), we model the attention state d_k at time instance k as a binary random variable, i.e.,

$$\begin{cases} d_k = 1, & \text{attending to speaker 1 at time instance } k \\ d_k = 2, & \text{attending to speaker 2 at time instance } k \end{cases}, \tag{5.33}$$

which is assumed to follow a Bernoulli distribution with probability p_k. The probability of attending to speaker 1 is given by

$$p_k = P(d_k = 1) = 1 - P(d_k = 2) = \frac{1}{1 + e^{-z_k}}, \tag{5.34}$$

where the variable z_k is modeled as an autoregressive (AR) process, i.e.,

$$z_k = c_0 z_{k-1} + w_k. \tag{5.35}$$

The parameter c_0 is a hyperparameter ensuring stability of the AR process and the noise process w_k is assumed to follow a normal distribution with variance η_k, i.e.,

$$w_k \sim \mathcal{N}(0, \eta_k), \tag{5.36}$$

$$\eta_k \sim \text{Inverse-Gamma}(a_0, b_0), \tag{5.37}$$

where a_0 and b_0 are hyperparameters controlling the smoothness of the SSM. The AR model in (5.35) implies that the z_k at time instance k is predicted from z_{k-1} at the previous time instance with some uncertainty, which is modeled by the noise process w_k.

To relate the correlation coefficients $\rho_{1,k}$ and $\rho_{2,k}$ in (5.9) to the attention state d_k, we model the probability of the absolute values of the correlation coefficients $\phi_{\{1,2\},k}$ given attention to speaker 1 or speaker 2, using a log-normal distribution[3], i.e.,

$$p(\phi_{l,k} \mid d_k = l) \sim \text{Log-Normal}(\delta_a, \mu_a), \quad l = 1, 2, \tag{5.38}$$

with

$$\delta_a \sim \text{Gamma}(\bar{\gamma}_a, \bar{\nu}_a), \;\; p(\mu_a \mid \delta_a) \sim \mathcal{N}(\bar{\mu}_a, \delta_a), \tag{5.39}$$

where $\bar{\gamma}_a$, $\bar{\nu}_a$ and $\bar{\mu}_a$ denote the hyperparameters of the attended log-normal distribution. Similarly, we model the probability of the absolute values of the correlation coefficients $\phi_{\{1,2\},k}$ given no attention to speaker 1 or speaker 2, as

$$p(\phi_{l,k} \mid d_k \neq l) \sim \text{Log-Normal}(\delta_u, \mu_u), \quad l = 1, 2, \tag{5.40}$$

with

$$\delta_u \sim \text{Gamma}(\bar{\gamma}_u, \bar{\nu}_u), \;\; p(\mu_u \mid \delta_u) \sim \mathcal{N}(\bar{\mu}_u, \delta_u), \tag{5.41}$$

where $\bar{\gamma}_u$, $\bar{\nu}_u$ and $\bar{\mu}_u$ denote the hyperparameters of the unattended log-normal distribution. Since a small overlap between the attended and the unattended log-normal distributions is desired for a reliable decoding performance, the hyperparameters $\bar{\gamma}_{\{a,u\}}$, $\bar{\nu}_{\{a,u\}}$ and $\bar{\mu}_{\{a,u\}}$ are tuned to minimize the overlap.

3 Please note that modeling the probabilities of the absolute values of the correlation coefficients with log-normal distributions allows for a closed-form iterative solution [24] compared to when modeling the probabilities of the correlation coefficients either with normal or von Mises-Fisher distributions [197].

Aiming at estimating the probability of attending to speaker 1 and speaker 2 at time instance $k = k^*$ (see Fig. 5.3), we now consider using the absolute values of the correlation coefficients within a sliding window of length $K_{\text{SSM}} = K_P + K_A + 1$, with K_P and K_A denoting the number of correlation coefficients prior to and after k^*, respectively. The set of parameters to be estimated in this window is given by $\mathbf{\Omega} = \{z_{k^*-K_P:k^*+K_A}, \eta_{k^*-K_P:k^*+K_A}, \delta_a, \mu_a, \delta_u, \mu_u\}$. This set can be estimated by maximizing the log-posterior function, i.e.,

$$\hat{\mathbf{\Omega}} = \arg\max_{\mathbf{\Omega}} \ell\left(\mathbf{\Omega} | \phi_{1,k}, \phi_{2,k}, k = k^* - K_P : k^* + K_A\right), \tag{5.42}$$

which is a maximum a posteriori (MAP) estimate and can be computed iteratively using the Expectation Maximization (EM) algorithm as in [24, 197].

Based on the estimated variable z_k the probability $p_k = P(d_k = 1)$ of attending to speaker 1 is obtained using (5.34). These probabilities are segmented into non-overlapping windows of length K_{NOL}, i.e.,

$$\mathbf{p}_i = \left[p_{(i-1)K_{\text{NOL}}+1}\ p_{(i-1)K_{\text{NOL}}+2}\ \cdots\ p_{iK_{\text{NOL}}}\right]^T. \tag{5.43}$$

The probability of attending to speaker 1 in window i is then computed as the mean of the probabilities, i.e.,

$$P\left(\hat{d}_i = 1\right) = \frac{1}{K_{\text{NOL}}} \sum_{n=1}^{K_{\text{NOL}}} p_{(i-1)K_{\text{NOL}}+n}, \tag{5.44}$$

with $\hat{d}_i$ the attention state in window i. The probability of attending to speaker 2 in window i is computed as

$$P\left(\hat{d}_i = 2\right) = 1 - P\left(\hat{d}_i = 1\right). \tag{5.45}$$

Based on these probabilities, it is decided that the participant attended to speaker 1 if $P\left(\hat{d}_i = 1\right) > P\left(\hat{d}_i = 2\right)$ or attended to speaker 2 otherwise. The probabilistic attention measure of the attended speaker $\hat{p}_{a,i}$ in window i is hence determined as

$$\begin{cases} \hat{p}_{a,i} = P\left(\hat{d}_i = 1\right), & \text{if } P\left(\hat{d}_i = 1\right) > P\left(\hat{d}_i = 2\right) \\ \hat{p}_{a,i} = P\left(\hat{d}_i = 2\right), & \text{otherwise.} \end{cases} \tag{5.46}$$

The probabilistic attention measure of the unattended speaker $\hat{p}_{u,i}$ is determined as $\hat{p}_{u,i} = 1 - \hat{p}_{a,i}$. The process flow of AAD using the SSM is depicted in Fig. 5.3.

The AAD algorithm using the SSM was implemented and run using MATLAB (MATLAB 1 of RAAD in Fig. 5.2). The hyperparameters in (5.35) and (5.37) were set to $c_0 = 1$, $a_0 = 2.008$ and $b_0 = 0.2016$, similarly as in [24]. The hyperparameters $\bar{\gamma}_a$, $\bar{\nu}_a$ and $\bar{\mu}_a$ in (5.39) were set by fitting a gamma and a normal

distribution to the absolute values of the correlation coefficients of the (oracle) attended speaker obtained from the calibration phase. Similarly, the hyperparameters $\bar{\gamma}_u$, $\bar{\nu}_u$ and $\bar{\mu}_u$ in (5.41) were set by fitting a gamma and a normal distribution to the absolute values of the correlation coefficients of the (oracle) unattended speaker obtained from the calibration phase. The SSM parameter set $\mathbf{\Omega}$ was estimated using the EM algorithm as in [24] with 20 iterations. For the LW–SSM algorithm using the large overlapping correlation window (i.e., $K_{\mathrm{COR}} = 960$ samples, 1 sample shift), a small SSM window of length $K_{\mathrm{SSM}} = 1$ sample (corresponding to $\frac{1}{64}$ seconds) with $K_P = 0$ and $K_A = 0$ was used. For the SW–SSM algorithm using the small non-overlapping correlation window (i.e., $K_{\mathrm{COR}} = 16$ samples), a large SSM window of length $K_{\mathrm{SSM}} = 60$ samples (corresponding to 15 seconds) with $K_P = 53$ (corresponding to 13.25 seconds) and $K_A = 6$ (corresponding to 1.50 seconds) was used as in [24]. The length of the window K_{NOL} in (5.43) was set such that both algorithms generated the probabilistic attention measure of the attended speaker $\hat{p}_{a,i}$ in (5.46) every 0.25 seconds. This means that for the LW–SSM algorithm a window of length $K_{\mathrm{NOL}} = 16$ samples was used, while for the SW–SSM algorithm a window of length $K_{\mathrm{NOL}} = 1$ sample was used. Each participant's own data were used for hyperparameter and parameter setting as well as for decoding. To evaluate the performance of the proposed LW–SSM and SW–SSM algorithms, the decoding performance for each participant was computed as the percentage of correctly decoded NOL windows.

5.2.5.2 *Adaptive gain controller (AGC)*

The probabilistic attention measure of the attended speaker $\hat{p}_{a,i}$ in window i either obtained using the GLM in (5.30) or using the SSM in (5.46), is then used to drive the AGC (see Fig. 5.2).

The speech signal $s_{1,i}$ of speaker 1 and the speech signal $s_{2,i}$ of speaker 2 are first segmented into non-overlapping windows of length K_{AGC}, i.e., for window i

$$\mathbf{s}_{1,i} = \begin{bmatrix} s_{1,(i-1)K_{\mathrm{AGC}}+1} & s_{1,(i-1)K_{\mathrm{AGC}}+2} & \cdots & s_{1,iK_{\mathrm{AGC}}} \end{bmatrix}^T, \tag{5.47}$$

$$\mathbf{s}_{2,i} = \begin{bmatrix} s_{2,(i-1)K_{\mathrm{AGC}}+1} & s_{2,(i-1)K_{\mathrm{AGC}}+2} & \cdots & s_{2,iK_{\mathrm{AGC}}} \end{bmatrix}^T. \tag{5.48}$$

Based on the AAD results for window i, the attended speech vector $\hat{\mathbf{s}}_{a,i}$ and the unattended speech vector $\hat{\mathbf{s}}_{u,i}$ are determined as

$$\begin{cases} \hat{\mathbf{s}}_{a,i} = \mathbf{s}_{1,i}, \ \hat{\mathbf{s}}_{u,i} = \mathbf{s}_{2,i} & \text{if the identified attended speaker is speaker 1} \\ \hat{\mathbf{s}}_{a,i} = \mathbf{s}_{2,i}, \ \hat{\mathbf{s}}_{u,i} = \mathbf{s}_{1,i} & \text{otherwise.} \end{cases} \tag{5.49}$$

By multiplying the attended speech vector $\hat{\mathbf{s}}_{a,i}$ with the gain $\lambda_{a,i}$ and multiplying the unattended speech vector $\hat{\mathbf{s}}_{u,i}$ with the gain $\lambda_{u,i}$, the objective of the AGC is to achieve a desired signal-to-interference-ratio (SIR) between the identified attended

and unattended speakers in window i. The desired SIR in window i is defined as a linear function of the probabilistic attention measure $\hat{p}_{a,i}$, i.e.,

$$\mathrm{SIR}_i^{des} = 2\mathrm{SIR}_{max}\hat{p}_{a,i} - \mathrm{SIR}_{max}, \tag{5.50}$$

such that $\hat{p}_{a,i} = 1$ corresponds to SIR_{max}, i.e., the maximum desired SIR and a $\hat{p}_{a,i} = 0.5$ corresponds to SIR = 0 dB. The SIR in window i at the output of the AGC is equal to

$$\mathrm{SIR}_i = 10\log_{10}(\frac{\lambda_{a,i}^2\varphi_{a,i}}{\lambda_{u,i}^2\varphi_{u,i}}), \tag{5.51}$$

with the energy of the attended and unattended speech signal in window i given by

$$\varphi_{a,i} = \hat{\mathbf{s}}_{a,i}^T\hat{\mathbf{s}}_{a,i}, \;\; \varphi_{u,i} = \hat{\mathbf{s}}_{u,i}^T\hat{\mathbf{s}}_{u,i}. \tag{5.52}$$

By setting (5.51) equal to the desired SIR in (5.50) and constraining the overall energy at the output of the AGC to be equal to the overall input energy, i.e.,

$$\lambda_{a,i}^2\varphi_{a,i} + \lambda_{u,i}^2\varphi_{u,i} = \varphi_{a,i} + \varphi_{u,i}, \tag{5.53}$$

the gains $\lambda_{u,i}$ and $\lambda_{a,i}$ can be computed as

$$\lambda_{u,i}^2 = \frac{1+\frac{\varphi_{a,i}}{\varphi_{u,i}}}{1+10^{\frac{\mathrm{SIR}_i^{des}}{10}}}, \tag{5.54}$$

$$\lambda_{a,i}^2 = 10^{\frac{\mathrm{SIR}_i^{des}}{10}} + \frac{\varphi_{u,i}}{\varphi_{a,i}}\lambda_{u,i}^2. \tag{5.55}$$

To avoid annoying artefacts due to highly time-varying gains, the gains $\lambda_{u,i}$ in (5.54) and $\lambda_{a,i}$ (5.55) are averaged over four windows, i.e.,

$$\bar{\lambda}_{u,i} = \frac{1}{4}\sum_{n=i-3}^{i}\lambda_{u,n}, \;\; \bar{\lambda}_{a,i} = \frac{1}{4}\sum_{n=i-3}^{i}\lambda_{a,n}. \tag{5.56}$$

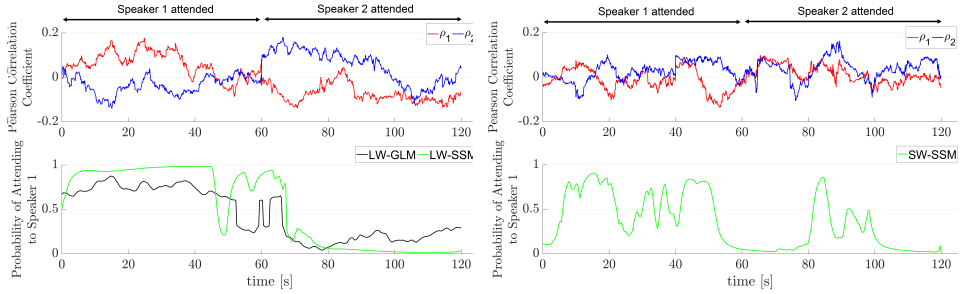
The amplified attended speaker $\bar{\boldsymbol{s}}_{a,i}$ and the attenuated unattended speaker $\bar{\boldsymbol{s}}_{u,i}$ in window i are finally obtained as

$$\bar{\boldsymbol{s}}_{a,i} = \bar{\lambda}_{a,i}\hat{\mathbf{s}}_{a,i}, \tag{5.57}$$

$$\bar{\boldsymbol{s}}_{u,i} = \bar{\lambda}_{u,i}\hat{\mathbf{s}}_{u,i}. \tag{5.58}$$

These signals are then presented to the participant using two loudspeakers. The AGC was implemented and run using MATLAB (MATLAB 2 in Fig. 5.2). The sampling frequency of the speech signals of both speakers was equal to 44100 Hz. The maximum desired SIR in (5.50) was set to +7. The speech enhancement performance of AGC was evaluated in terms of the SIR improvement ΔSIR, i.e.,

$$\Delta\mathrm{SIR} = \mathrm{SIR}_{out} - \mathrm{SIR}_{in}, \tag{5.59}$$



(a) Large correlation window ($K_{\text{COR}} = 960$ samples, 1 sample shift)

(b) Small correlation window ($K_{\text{COR}} = 16$ samples, no overlap)

Fig. 5.4: Exemplary correlation coefficients of speaker 1 and speaker 2 and probabilistic attention measures of speaker 1 from the open-loop AAD phase when using AAD algorithms employing (a) a large correlation window (LW–GLM, LW–SSM), and (b) a small correlation window (SW–SSM).

with

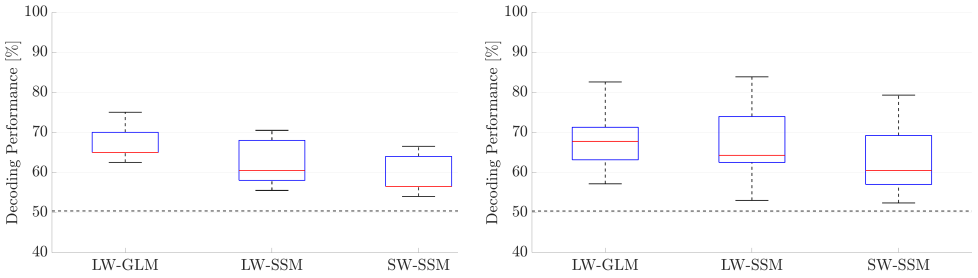
$$\text{SIR}_{in} = 10\log_{10}\left(\frac{\sum_{i=1}^{I} \boldsymbol{s}_{a,i}^T \boldsymbol{s}_{a,i}}{\sum_{i=1}^{I} \boldsymbol{s}_{u,i}^T \boldsymbol{s}_{u,i}}\right), \tag{5.60}$$

$$\text{SIR}_{out} = 10\log_{10}\left(\frac{\sum_{i=1}^{I} \bar{\boldsymbol{s}}_{a,i}^T \bar{\boldsymbol{s}}_{a,i}}{\sum_{i=1}^{I} \bar{\boldsymbol{s}}_{u,i}^T \bar{\boldsymbol{s}}_{u,i}}\right), \tag{5.61}$$

where $\boldsymbol{s}_{a,i}$ and $\boldsymbol{s}_{u,i}$ denote the (oracle) attended and unattended speech vectors, defined similarly as in (5.47).

5.3 Results

In this section, we evaluate the decoding performance and the speech enhancement performance of the proposed cognitive-driven gain controller system described in the previous section. In Section 5.3.1 we evaluate the decoding performance of the proposed AAD algorithms for the open-loop and the closed-loop AAD phase. In Section 5.3.2 we evaluate the speech enhancement performance of the AGC for the closed-loop AAD phase. Finally, in Section 5.3.3 we compare the subjective evaluation between the open-loop and the closed-loop AAD phase.



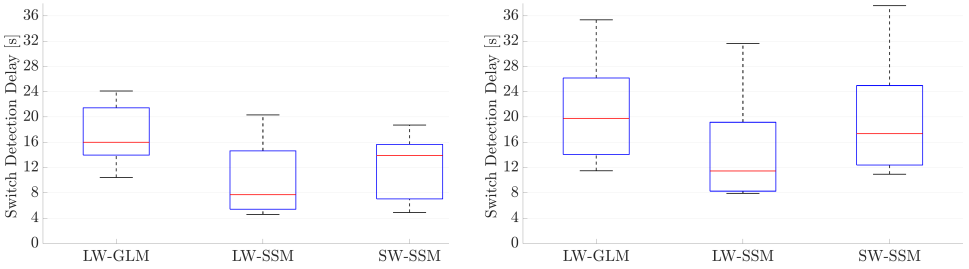
(a) open-loop AAD (b) closed-loop AAD

Fig. 5.5: Decoding performance for (a) the open-loop AAD and (b) the closed-loop AAD when using the LW–GLM, the LW–SSM and the SW–SSM. The dashed-line represents the upper boundary of the confidence interval corresponding to chance level based on a binomial test at the 5% significance level.

5.3.1 *Auditory attention decoding performance*

For the AAD algorithms using the large correlation window (LW–GLM, LW–SSM) and the small correlation window (SW–SSM), Fig. 5.4 depicts exemplary correlation coefficients $\rho_{1,k}$ and $\rho_{2,k}$ of speaker 1 and speaker 2, corresponding to a session from the open-loop AAD phase. It can be observed that all considered AAD algorithms translate the correlation coefficients into smooth probabilistic attention measures, which are robust against fluctuations of the correlation coefficients. When using the large correlation window, i.e., LW–GLM and LW–SSM, the correlation coefficients are more discriminative and the probabilistic attention measures are more reliable while having a low variability compared to using the small correlation window, i.e., SW–SSM. This can mainly be explained by the fact that the large correlation window provides a larger amount of data from the reconstructed attended envelope and the envelopes of the speakers compared to the small correlation window. The discriminability and reliability of correlation coefficients and probabilistic attention measures are obviously essential ingredients to improve the decoding performance.

For the considered AAD algorithms, Fig. 5.5 depicts the decoding performance for the open-loop and the closed-loop AAD phase. It can be observed that all AAD algorithms yield a decoding performance that is larger than chance level (50.0%). For the open-loop AAD phase, the LW–GLM, LW–SSM and SW–SSM algorithms yield a (median) decoding performance of 65.0%, 60.5% and 56.5%, respectively. For the closed-loop AAD phase, the LW–GLM, LW–SSM and SW–SSM algorithms yield a (median) decoding performance of 67.7%, 64.2% and 60.4%, respectively. The higher median decoding performance obtained by the LW–GLM and LW–SSM algorithms is consistent with the probabilistic attention measures in Fig. 5.4, where due to the large correlation window more reliable probabilistic attention measures are obtained compared to the SW–SSM algorithm. In order to compare the decoding performance between the open-loop and the closed-loop AAD phase, a statistical multiple comparison test (Kruskal-Wallis test followed by the post-hoc Dunn and



(a) open-loop AAD (b) closed-loop AAD

Fig. 5.6: Delay to detect a cued attention switch for (a) the open-loop AAD phase and (b) the closed-loop AAD phase when using the LW–GLM, LW–SSM and SW–SSM algorithms.

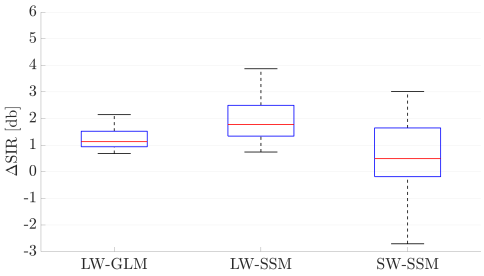


Fig. 5.7: SIR improvement of the proposed cognitive-driven gain controller system when using the LW–GLM, LW–SSM and SW–SSM algorithms.

Sidak test [1]) was performed. This test revealed no significant difference ($p > 0.05$) in decoding performance between the open-loop and the closed-loop AAD phase nor the considered AAD algorithms.

To further investigate the performance of the proposed AAD algorithms, Fig. 5.6 depicts the delay to detect a cued attention switch for the open-loop and the closed-loop AAD phase. For the open-loop AAD phase, the LW–GLM, LW–SSM and SW–SSM algorithms yield a median delay of 16.0 seconds, 7.7 seconds and 13.9 seconds, respectively. For the closed-loop AAD phase, the LW–GLM, LW–SSM and SW–SSM algorithms yield a median delay of 19.8 seconds, 11.5 seconds and 17.4 seconds, respectively. A statistical multiple comparison test (Kruskal-Wallis test followed by the post-hoc Dunn and Sidak test) revealed no significant difference ($p > 0.05$) between the open-loop and the closed-loop AAD phase nor the considered AAD algorithms.

5.3.2 *Speech enhancement performance of adaptive gain controller*

For the considered AAD algorithms, Fig. 5.7 depicts the SIR improvement for the closed-loop AAD phase. It can be observed that the LW–GLM, LW-GLM and SW–SSM algorithms yield a median SIR improvement of 1.1 dB, 1.7 dB and 0.5 dB, respectively. The larger SIR improvement obtained by the LW–GLM and LW–SSM algorithms can be explained by the higher decoding performance compared to the SW–SSM algorithm. The higher decoding performance leads to a larger number of windows during which the attended speaker is correctly amplified and the unattended speaker is correctly attenuated. In addition, it can be observed that the SW-SSM algorithm yields an SIR improvement with a larger variability (-2.7-3.0 dB) than the LW–GLM algorithm (0.6-2.1 dB) and the LW–SSM algorithm (0.7-3.8 dB). This can be explained by the larger variability of the probabilistic attention measures obtained by the SW–SSM (see Fig. 5.4). Due to the linear role of the probabilistic attention measure in the AGC for determining the desired SIR between the attended and the unattended speaker, see (5.50), the probabilistic attention measures with a large variability lead to SIRs with a large variability.

5.3.3 *Subjective evaluation of open-loop and closed-loop AAD*

Finally, we subjectively evaluate the open-loop and the closed-loop AAD phase. Fig. 5.8 presents the perceived effort to follow the attended speaker, to ignore the unattended speaker, to switch attention between both speakers, and the level of story understanding.

For the effort to follow the attended speaker and to ignore the unattended speaker (Fig. 5.8a and Fig. 5.8b), it can be observed that the lowest median effort level is obtained for the open-loop AAD, while a higher median is obtained for the closed-loop AAD, especially when using the SW–SSM algorithm. This can be attributed to the negative SIR improvements in some windows (see Fig. 5.7), where the attended speaker is wrongly attenuated and the unattended speaker is wrongly amplified. Nevertheless, a statistical multiple comparison test (Kruskal-Wallis test followed by the post-hoc Dunn and Sidak test) revealed no significant difference ($p > 0.05$) between all considered open-loop and closed-loop AAD cases. Similarly, for the effort to switch attention between both speakers (Fig. 5.8c), a statistical multiple comparison test revealed no significant difference ($p > 0.05$) between all considered open-loop and closed-loop AAD cases. These results show that the proposed closed-loop cognitive-driven gain controller system demands a similar perceived effort to follow the attended speaker, to ignore the unattended speaker and to switch attention compared to the open-loop AAD system. For the level of story understanding (Fig. 5.8d), the highest median understanding level is obtained for the open-loop AAD, while a lower median understanding level is obtained for the closed-loop AAD. This is consistent with the perceived level of cognitive effort (Fig. 5.8a, Fig. 5.8b, Fig. 5.8c), where the open-loop AAD demands the lowest median effort level, possibly resulting in more cognitive resources available for story understanding compared to the closed-loop AAD. Nevertheless, a statistical multiple comparison test

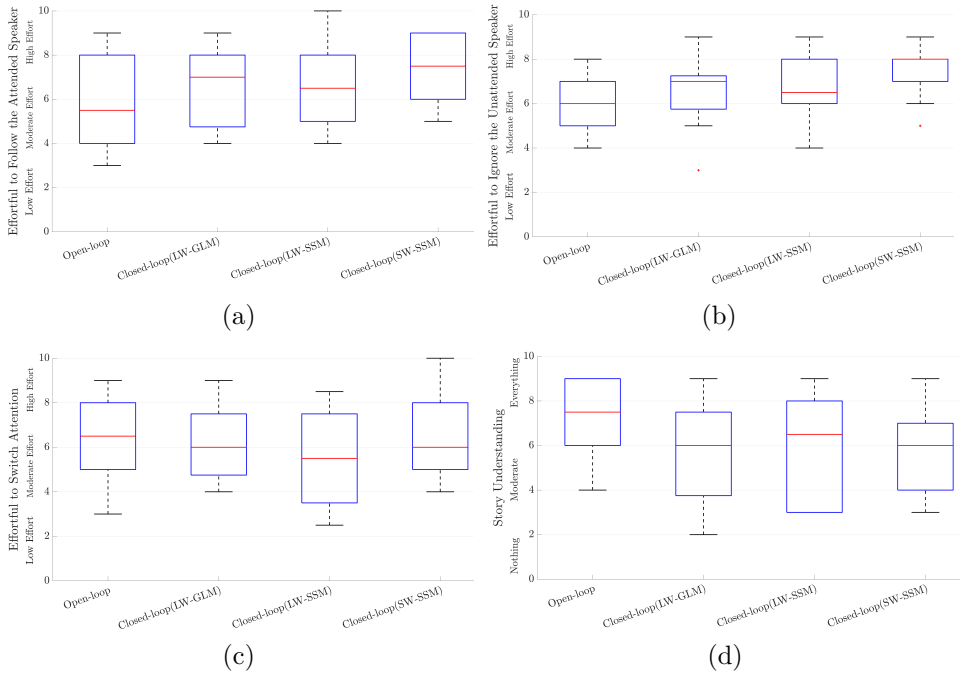


Fig. 5.8: Subjective evaluation results of the open-loop and the closed-loop AAD phase using the LW–GLM, LW–SSM and SW–SSM algorithms in terms of perceived effort to (a) follow the attended speaker, (b) ignore the unattended speaker, (c) switch attention between both speakers and (d) understand the story.

revealed no significant difference ($p > 0.05$) between the considered open-loop and closed-loop AAD cases.

Lastly, Fig. 5.9 presents the level of improvement in system usage achieved by the participants throughout the sessions of closed-loop AAD experiment. It can be observed for all considered AAD algorithms that a significant improvement in system usage is obtained.

5.4 Discussion

The experimental results for the open-loop AAD system show that the highest median decoding performance is obtained by the LW–GLM algorithm (65%). This is in accordance with the experimental results in [20], where it has been shown that open-loop AAD using a low number of electrodes with a correlation window smaller

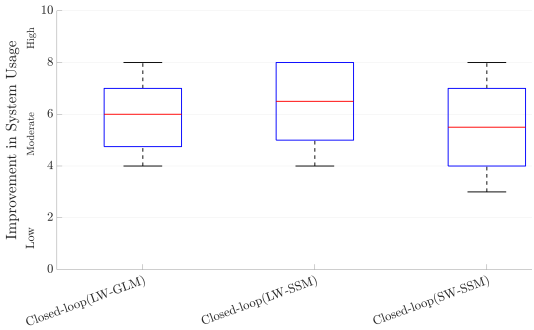


Fig. 5.9: Subjective improvement in system usage for the closed-loop AAD system when using the LW–GLM, LW–SSM and SW–SSM algorithms.

than 15 seconds results in a decoding performance lower than 75%[4]. In addition, the experimental results show that there is no significant difference in decoding performance between the open-loop and the closed-loop AAD system using the proposed AAD algorithms. This is consistent with the experimental results in [163], where no significant difference in decoding performance between an open-loop and a closed-loop AAD system using visual feedback has been observed.

The experimental results show that the LW–GLM and LW–SSM algorithms using the large correlation window yield a higher median decoding performance compared to the SW–SSM algorithm using the small correlation window. The large correlation window provides a larger amount of data from the reconstructed attended envelope and the envelopes of the speakers compared to the small correlation window, resulting in more discriminative correlation coefficients, more reliable probabilistic attention measures and a larger decoding performance. This is in accordance with the experimental results in [18, 20], where it has been shown that using a larger correlation window results in a larger decoding performance. The experimental results also show that the LW–GLM algorithm yields a higher median decoding performance compared to the LW–SSM algorithm. This may be explained by the fact that the LW–GLM algorithm infers the probabilistic attention measures based on the mean differential correlation coefficients rather than the absolute values of the correlation coefficient. The mean differential correlation coefficients provide a larger dynamic range including negative and non-negative values for inferring the probabilistic attention measures.

The experimental results demonstrate the feasibility of closed-loop AAD in an online fashion, enabling the listener to interact with an adaptive gain controller (as an ideal speech enhancement algorithm) for a two-speaker scenario. On the one hand, the closed-loop cognitive-driven gain controller system improves the SIR between

4 Please note that the decoding performance in [20] was obtained based on an optimal EEG electrode configuration, whereas the decoding performance reported in this paper was obtained based on a fixed EEG electrode configuration.

the attended and the unattended speaker. This may make it easier to follow the attended speaker, ignore the unattended speaker and switch attention between both speakers, resulting in a lower cognitive effort, compared to open-loop AAD. On the other hand, the closed-loop cognitive-driven gain controller system introduces a significant delay to detect attention switches, which causes the attended speaker to be wrongly attenuated and the unattended speaker to be wrongly amplified. This may make it more difficult to follow the attended speaker and ignore the unattended speaker, resulting in a higher cognitive effort compared to open-loop AAD. Nevertheless, the subjective evaluation results indicate that overall the closed-loop cognitive-driven gain controller system demands a similar effort as the open-loop AAD system.

The latency of the proposed closed-loop cognitive-driven gain controller system affecting the processing of the attended and unattended speech signals consists of three parts: correlation coefficient generation, AAD using either the large-window GLM, the large-window SSM or the small-window SSM, and adaptive gain controller. The latency caused by the correlation coefficient generation using the large correlation window and the small correlation window correspond to $\frac{1}{64}$ seconds and $\frac{1}{4}$ seconds, respectively. The latency caused by AAD using the large-window GLM corresponds to $\frac{1}{4}$ seconds. The latency caused by AAD using the large-window SSM (SSM window with $K_A = 0$) corresponds to $\frac{1}{64}$ seconds, whereas the latency caused by AAD using the small-window SSM (SSM window with $K_A = 6$) corresponds to 1.50 seconds. Also, the latency for generating the probabilistic attention measures caused by AAD using either the large-window SSM or the small-window SSM corresponds to $\frac{1}{4}$ seconds. The latency caused by the adaptive gain controller corresponds to $\frac{1}{44100}$ seconds.

While the closed-loop AAD experiments using the proposed cognitive-driven gain controller system was performed without incorporating a practicing phase for participants in this paper, the subjective evaluation results suggest that a significant improvement in the system usage was obtained throughout the closed-loop AAD experiment. Future work could therefore investigate the impact of incorporating a practicing phase on the decoding and the speech enhancement performance of the cognitive-driven gain controller system.

5.5 Conclusion

In this paper, we proposed a closed-loop gain controller system which cognitively steers an adaptive gain controller based on real-time AAD for a scenario with two competing speakers. The real-time AAD infers the probabilistic attention measures of the attended and the unattended speaker from EEG recordings of the listener and the speech signals of both speakers. Based on these probabilistic attention measures, the adaptive gain controller amplifies the identified attended speaker and attenuates the identified unattended speaker. The loop of cognitive-driven gain control is then closed by presenting the amplified attended speaker and the attenuated unattended speaker via loudspeakers. The experimental results demonstrate the feasibility of the proposed closed-loop cognitive-driven gain controller system (both using GLM and

SSM), enabling the listener to interact with the system in real-time. Although there is a significant delay to detect attention switches, which causes the attended speaker to be wrongly attenuated and the unattended speaker to be wrongly amplified, the proposed closed-loop system is able to improve the SIR between the attended and the unattended speaker. Moreover, the subjective evaluation results show that the proposed closed-loop cognitive-driven system demands a similar perceived level of cognitive effort to follow the attended speaker, to ignore the unattended speaker and to switch attention between both speakers compared to the open-loop AAD system. With this work, a first attempt was made to bring closed-loop cognitive-driven speech enhancement closer to real-world applications.

6

EEG-BASED AUDITORY ATTENTION DECODING USING BINARY MASKING

The performance of many speech enhancement algorithms in a multi-speaker scenario depends on correctly identifying the target speaker to be enhanced. Recent advances in electroencephalography (EEG) have shown that it is possible to identify the target speaker from single-trial EEG recordings using a least-squares-based auditory attention decoding method. Since in practice the clean speech signals of the speakers are typically not available as reference signals for decoding, it has been proposed to either just use the (noisy and reverberant) microphone signals as reference signals or to generate reference signals by applying minimum-variance-distortionless-response (MVDR) beamformers to the microphone signals. However, since the MVDR output signals still contain interfering speech and background noise, the decoding performance is significantly lower compared to using the clean speech signals as reference signals. Aiming at generating better reference signals for decoding, in this paper we propose reference signal generation algorithms, which use binary masks to discard intervals with a low target speech energy, which are susceptible to interfering speech and background noise. The binary masks are determined from the directional speech presence probability of the speakers which can be estimated from the microphone signals. As reference signals we either use the masked microphone signals, the masked MVDR output signals or the binary masks themselves. The performance of the proposed algorithms is evaluated in terms of the decoding performance for a two-speaker scenario with a binaural hearing aid setup in anechoic and reverberant conditions. In addition, the performance of the proposed algorithms is analyzed based on the correlation difference and the signal-to-interference-plus-noise ratio (SINR). The experimental results show that using the masked microphone and MVDR output signals as reference signals yields a larger correlation difference, SINR and decoding performance (especially in the reverberant condition) compared to using the (non-masked) microphone and MVDR output signals. Quite remarkably, the results also show that using the binary masks as reference signals yields a decoding performance that is comparable to using the masked MVDR output signals.

6.1 Introduction

In acoustic scenarios with multiple speakers and background noise the human auditory system has a remarkable ability to localize and segregate a speaker of interest and suppress the other speakers and background noise [45, 217]. For persons with hearing impairment source localization and segregation is typically much more challenging [218]. In order to improve speech intelligibility, several multi-microphone speech enhancement algorithms for hearing aid are currently available to perform source separation in multi-speaker scenario and to reduce background noise [10, 13, 115]. These algorithms aim at enhancing the target speaker, which is typically identified as the speaker in front of the hearing aid user or as the loudest speaker. Since in real-world conditions these assumptions about the target speaker may not correspond with the hearing wish of the listener, the benefit from speech enhancement algorithms may substantially decrease. Therefore, correctly identifying the target speaker in hearing aid applications is an essential ingredient to improve speech intelligibility.

Several studies have shown that auditory EEG responses of the listener are correlated with the envelope of the attended speaker [16,175,178]. Using single-trial EEG recordings, several auditory attention decoding (AAD) methods have been proposed to identify the attended speaker based, e.g., on a least-squares cost function [16], neural networks [23,130], and Bayesian filtering [24,172]. The possibility of decoding auditory attention from EEG recordings has led to the idea of incorporating AAD in a brain-computer interface, e.g., to control a hearing aid [198]. Hence, a large research effort has recently focused on investigating the feasibility of AAD in real-world listening conditions [20, 24, 26, 169, 198] and on steering speech enhancement algorithms based on AAD [27–29, 31, 166, 212].

The frequently used least-squares-based AAD method proposed in [16] aims at reconstructing the attended speech envelope from the EEG recordings using a trained spatio-temporal filter. The reconstructed envelope is then correlated with the envelope of two (attended and unattended) reference signals to identify the attended speaker. In [16] the the clean speech signals of both the attended and the unattended speaker have been used as reference signals for decoding. In practice, these clean speech signals are typically not available, i.e., only the microphone signals, containing interfering speech, background noise and reverberation, are available. Although in [26,169] it has been shown that AAD is still feasible using the microphone signals instead of the clean speech signals as reference signals, the decoding performance is significantly decreased, such that several source separation and noise reduction algorithms have been proposed to generate appropriate reference signals from the microphone signals [27–29, 31, 166, 212]. The single-microphone source separation algorithms proposed in [28,212] use deep neural networks to generate the reference signals by separating the speakers from the mixture received at the microphone. Although these algorithms are able to generate appropriate reference signals in anechoic conditions, the generalization to reverberant conditions is still challenging. The multi-microphone algorithm proposed in [27, 166] generates reference signals using multiple multi-microphone Wiener filters (MWFs). These MWFs rely on the second-order statistics of the interfering speech and background noise, which are estimated

using an envelope demixing algorithm and a voice activity detector. Experimental results in [27,166] show that using the MWF-based reference signals yields a larger decoding performance (especially in an anechoic condition) than using the microphone signals as reference signals. However, the performance of the MWFs highly relies on the estimation accuracy of the second-order statistics provided by the envelope demixing algorithm, which is prone to reverberation. Instead of using multiple MWFs, the multi-microphone algorithm proposed in [29,31] generates reference signals using multiple minimum-variance-distortionless-response (MVDR) beamformers, which are steered based on the estimated directional-of-arrivals (DOAs) of the speakers. Although experimental results in [29,31] show that using the MVDR output signals as reference signals yields a larger decoding performance than using the microphone signals as reference signals, the decoding performance is significantly lower compared to using the clean speech signals as reference signals. This can be explained by the fact that the interfering speaker may not be sufficiently suppressed by the MVDR beamformers, leading to intervals in the reference signals where the interfering speaker (and the background noise) dominate, e.g., when the target speaker pauses or speaks softly.

Aiming at generating better reference signals for decoding, in this paper we propose a reference signal generation algorithm, which uses binary masks to discard (low-energy) intervals which are susceptible to interfering speech and background noise (see Fig. 6.1). First, the directional speech presence probability (DSPP) and the DOA of both speakers are estimated from the microphone signals. Based on the estimated DSPPs, the intervals with low estimated speech energy for both speakers are detected and binary masks are estimated. The reference signals for decoding are then generated by either masking the microphone signals or the MVDR output signals. In addition, the estimated binary masks are considered themselves as reference signals for decoding.

The paper is organized as follows. In Section 6.2 the configuration and notation used for the binaural hearing aid setup are introduced. In Section 6.3 the reference signal generation using MVDR beamformers is briefly reviewed, where also the algorithm to estimate the directional speech presence probability and the DOAs are described. In Section 6.4 the reference signal generation based on binary masking is described. In Section 6.5 the least-squares-based AAD method using the generated reference signals is briefly reviewed. In Section 6.7 the experimental results are presented, investigating the decoding performance of the proposed reference signal generation algorithms for an acoustic scenario comprising two competing speakers and diffuse background noise in an anechoic and a reverberant condition.

6.2 Configuration and notation

We consider an acoustic scenario comprising two competing speakers with DOAs θ_1 and θ_2 and background noise in a reverberant environment (see Fig. 6.1). The angle $\theta = 0^\circ$ corresponds to the frontal direction, with negative θ corresponding to the left side of the listener and positive θ corresponding to the right side. The clean signals of the speakers are denoted as $s_1[n]$ and $s_2[n]$, with n the discrete time

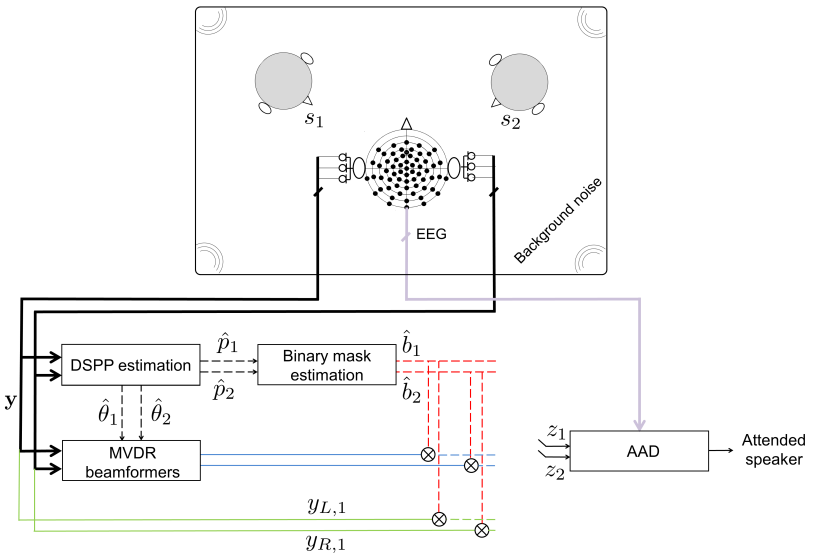


Fig. 6.1: Block diagram of the proposed reference signal generation algorithm in an acoustic scenario comprising two competing speakers (s_1 and s_2) on the left and the right side of the listener. First, the directional speech presence probability (DSPP) and the DOA of speakers are estimated from the microphone signals $\mathbf{y}$. Based on the estimated DSPPs ($\hat{p}_1$ and $\hat{p}_2$), the intervals with low estimated speech energy for both speakers are detected and binary masks ($\hat{b}_1$ and $\hat{b}_2$) are estimated. The reference signals (z_1 and z_2) are then generated by either masking the microphone signals or the output signals of the MVDR beamformers or bu considering the estimated binary masks themselves. Using the generated reference signals together with single-trial EEG recording of the listener, the attended speaker is then identified.

index. We consider a binaural hearing aid setup, where each hearing aid contains M microphones. The m-th microphone signal $y_{L,m}[n]$ of the left hearing aid can be decomposed as

$$y_{L,m}[n] = x_{1,L,m}[n] + x_{2,L,m}[n] + v_{L,m}[n], \tag{6.1}$$

where $x_{1,L,m}[n]$ and $x_{2,L,m}[n]$ denote the reverberant speech component in the m-th microphone signal corresponding to speaker 1 and speaker 2, respectively, and $v_{L,m}[n]$ denotes the background noise component. The reverberant speech components $x_{1,L,m}[n]$ and $x_{2,L,m}[n]$ consist of an anechoic speech component $x^{an}_{1,L,m}[n]$ and $x^{an}_{2,L,m}[n]$, encompassing the (anechoic) head filtering effect, and a reverberation component. The m-th microphone signal $y_{R,m}[n]$ of the right hearing aid can be decomposed similarly as in (6.1)

In the short-time Fourier transform (STFT) domain, the $2M$-dimensional stacked vector of all microphone signals from the left and the right hearing aid is given by

$$\mathbf{y}(k,l) = [Y_{L,1}(k,l) \ \dots \ Y_{L,M}(k,l) \ \ Y_{R,1}(k,l) \ \dots \ Y_{R,M}(k,l)]^T, \tag{6.2}$$

where k denotes the frequency index and l denotes the frame index. For notational conciseness the indices k and l will be omitted in the remainder of this paper wherever possible. Using (6.1) and (6.2), the signal vector $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{v}, \tag{6.3}$$

where the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{v}$ are defined similarly as in (6.2) for speaker 1, speaker 2, and the background noise component, respectively. To decode auditory attention from the listener, the least-squares-based AAD method (see Section 6.5) requires two reference signals $z_1[n]$ and $z_2[n]$, corresponding to speaker 1 and speaker 2.

In [16] the clean speech signals $s_1[n]$ and $s_2[n]$ have been used as reference signals. Since the anechoic speech components of speaker 1 and speaker 2 in the microphone signals encompassing the head filtering effect are more similar to what the listener perceives compared to the clean speech signals, in this paper (similarly as in [26]) the anechoic speech component in the first microphone signal at the side of speaker 1 and speaker 2 will be used as oracle reference signals, i.e.,

$$z_1[n] = x_{1,L,1}^{an}[n], \quad z_2[n] = x_{2,R,1}^{an}[n]. \tag{6.4}$$

Since these speech components are typically not available in practice, we will refer to the reference signals in (6.4) as oracle reference signals (ORACLE). Since due to the head shadowing effect it can be assumed for the considered scenario (see Fig. 6.1) that the broadband energy of the speech components $x_{1,L,m}[n]$ in the left microphone signals corresponding to speaker 1 is larger than the broadband energy of the speech components $x_{2,L,m}[n]$ corresponding to speaker 2 (and vice-versa for the right microphone signals), it has been shown in [32,169] that AAD is also feasible using the unprocessed microphone signals as reference signals, i.e.,

$$z_1[n] = y_{L,1}[n], \quad z_2[n] = y_{R,1}[n]. \tag{6.5}$$

We will refer to the reference signals in (6.5) as the microphone reference signals (MIC). Obviously, the decoding performance is heavily affected by the presence of background noise and especially the interfering speaker in the microphone reference signals.

In the following sections we will discuss several algorithms to generate better reference signals from the microphone signals. In Section 6.3 we will briefly review the reference signal generation algorithm using multiple MVDR beamformers [29, 31]. In Section 6.4 we will propose a novel reference signal generation algorithm based on directional speech presence probability and binary masking.

6.3 Reference signal generation using MVDR beamformers

In [29, 31] it has been proposed to generate the reference signals corresponding to speaker 1 and speaker 2 using two binaural MVDR beamformers, which are steered based on the estimated DOAs of both speakers. Section 6.3.1 describes an algorithm to estimate the directional speech presence probability and the DOAs

from the binaural microphone signals. Section 6.3.2 describes the binaural MVDR beamformers.

6.3.1 *DSPP and DOA estimation*

Several multi-microphone methods have been proposed in the literature to estimate directional speech presence probability and voice activity detection [90,199,219–222]. In this paper, we will use the discriminative learning approach from [90], where support vector machine (SVM) classifiers estimate the source presence probability for different DOAs. The SVMs are trained to distinguish between the presence of a source for a certain direction and the absence for all other directions. The decision value of each SVM is mapped to a directional source presence probability $p_\theta[n]$ for each direction using a generalized linear model. As feature the short-term generalized cross-correlation with phase transform (GCC-PHAT) function in the time-domain [200] is used, which has been shown to be relatively robust to reverberation. To increase the robustness against background noise, we first smooth the DSPP for each direction across time, i.e.,

$$\overline{p}_\theta[n] = \tau p_\theta[n] + (1-\tau)\,\overline{p}_\theta[n-1], \tag{6.6}$$

with τ denoting the recursive smoothing constant. We then select two DOAs with the largest smoothed DSPP $\overline{p}_\theta[n]$, from which the DOAs of speaker 1 and speaker 2 are determined such that $\hat{\theta}_1 \leq \hat{\theta}_2$. The DSPP of both speakers is then determined as

$$\hat{p}_1[n] = p_{\hat{\theta}_1}[n], \tag{6.7}$$

$$\hat{p}_2[n] = p_{\hat{\theta}_2}[n]. \tag{6.8}$$

6.3.2 *MVDR beamformer*

Aiming at generating better reference signals for decoding, it has been proposed in [29,31] to apply two MVDR beamformers to the binaural microphone signals. The MVDR beamformer [9, 10, 86] aims at minimizing the noise power spectral density (PSD) while preserving sounds arriving from target direction θ_t. The corresponding constrained optimization problem is given by

$$\min_{\mathbf{w}} \underbrace{\mathbf{w}^H \mathbf{\Phi}_{\mathbf{v}} \mathbf{w}}_{\text{noise output PSD}} \quad \text{subject to } \underbrace{\mathbf{w}^H \mathbf{a}(\theta_t) = 1}_{\text{target}}, \tag{6.9}$$

where $\mathbf{w}$ denotes the 2M-dimensional filter vector applied to all microphone signals, $\mathbf{a}(\theta)$ denotes the (anechoic) steering vector for direction θ and $\mathbf{\Phi}_{\mathbf{v}} = E\{\mathbf{v}\mathbf{v}^H\}$ denotes the noise covariance matrix with $E\{\cdot\}$ the expected value operator. The MVDR beamformer solving (6.9) is given by

$$\mathbf{w}_{\text{MVDR}}(\theta_t) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}(\theta_t)}{\mathbf{a}^H(\theta_t)\,\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}(\theta_t)}. \tag{6.10}$$

To generate reference signals from the binaural microphone signals, two MVDR beamformers are used, i.e., an MVDR beamformer with estimated target angle $\theta_t = \hat{\theta}_1$ to generate the reference signal for speaker 1, and an MVDR beamformer with estimated target angle $\theta_t = \hat{\theta}_2$ to generate the reference signal for speaker 2, i.e.,

$$z_1[n] = \text{ISTFT}\left\{\mathbf{w}_{\text{MVDR}}^H(\hat{\theta}_1)\boldsymbol{y}\right\}, \tag{6.11}$$

$$z_2[n] = \text{ISTFT}\left\{\mathbf{w}_{\text{MVDR}}^H(\hat{\theta}_2)\boldsymbol{y}\right\}. \tag{6.12}$$

where ISTFT denotes the inverse short-time Fourier transform. We refer to the reference signals in (6.11) and (6.12) as the MVDR reference signals.

Although it has been shown in [29,31] that using the MVDR reference signals yields a larger decoding performance than using the microphone reference signals, the decoding performance is significantly lower compared to using the oracle reference signals. This can be explained by the fact that the MVDR beamformers may not be able to sufficiently suppress the interfering speaker, i.e., speaker 2 in $z_1[n]$ and speaker 1 in $z_2[n]$.

6.4 Reference signal generation using binary masking

In this section, we propose an algorithm to improve the decoding performance based on binary masking. The main idea is to discard intervals in the reference signals with a low target speech energy, which are hence susceptible to interfering speech and background noise. After defining oracle binary masks, Section 6.4.1 describes an approach to estimate the binary masks from DSPPs discussed in Section 6.3.1. Section 6.4.2 then discusses how to use the estimated binary masks to generate reference signals for decoding.

6.4.1 *Binary mask estimation*

To perfectly detect the intervals in the reference signals with low target speech energy, i.e., when the target speaker pauses or speaks softly, we require the target speech component in the reference signals. Assuming the oracle reference signals in (6.4) are available, the ideal binary masks for speaker 1 and speaker 2 are defined as

$$b_1[n] = \begin{cases} 0 & \text{if } \gamma_1[n] \leq \alpha_{1,\gamma} \\ 1 & \text{else} \end{cases}, \tag{6.13}$$

$$b_2[n] = \begin{cases} 0 & \text{if } \gamma_2[n] \leq \alpha_{2,\gamma} \\ 1 & \text{else} \end{cases}. \tag{6.14}$$

where $\gamma_1[n] = \left|x_{1,L,1}^{an}[n]\right|^2$ and $\gamma_2[n] = \left|x_{2,R,1}^{an}[n]\right|^2$ denote the energy of speaker 1 and speaker 2, and $\alpha_{1,\gamma}$ and $\alpha_{2,\gamma}$ denote the energy threshold for speaker 1 and speaker 2. Since in practice the oracle reference signals are not available, we propose to detect the susceptible intervals based on the estimated DSPPs $\hat{p}_1[n]$ and $\hat{p}_2[n]$

of the speakers, which can be estimated from the microphone signals (see Section 6.3.1). The binary masks for speaker 1 and speaker 2 are estimated as

$$\hat{b}_1[n] = \begin{cases} 0 & \text{if } \hat{p}_1[n] \leq \alpha_{1,p} \\ 1 & \text{else} \end{cases}, \tag{6.15}$$

$$\hat{b}_2[n] = \begin{cases} 0 & \text{if } \hat{p}_2[n] \leq \alpha_{2,p} \\ 1 & \text{else} \end{cases}, \tag{6.16}$$

where $\alpha_{1,p}$ and $\alpha_{2,p}$ denote the DSPP threshold for speaker 1 and speaker 2.

6.4.2 *Mask reference signals*

To generate reference signals without low target speech energy intervals, the ideal or estimated binary masks can be applied either to the microphone signals, i.e.,

$$z_1[n] = \hat{b}_1[n]\, y_{L,1}[n], \tag{6.17}$$

$$z_2[n] = \hat{b}_2[n]\, y_{R,1}[n]. \tag{6.18}$$

or to the MVDR output signals, i.e.,

$$z_1[n] = \hat{b}_1[n]\, \text{ISTFT}\left\{\mathbf{w}_{\text{MVDR}}^H\left(\hat{\theta}_1\right)\boldsymbol{y}\right\}, \tag{6.19}$$

$$z_2[n] = \hat{b}_2[n]\, \text{ISTFT}\left\{\mathbf{w}_{\text{MVDR}}^H\left(\hat{\theta}_2\right)\boldsymbol{y}\right\}. \tag{6.20}$$

In addition, the binary masks themselves may also be used directly as reference signals, i.e., $z_1[n] = b_1[n]$ and $z_2[n] = b_2[n]$ or $z_1[n] = \hat{b}_1[n]$ and $z_2[n] = \hat{b}_2[n]$.

6.5 Auditory attention decoding

Based on the reference signals z_1 and z_2 and the EEG recordings of the listener, auditory attention is then decoded. This section briefly reviews the least-squares-based AAD method proposed in [16], which aims at reconstructing the attended speech envelope from the EEG recordings using a trained spatio-temporal filter. Section 6.5.1 describes the decoding step, where the reconstructed envelope is compared with the envelope of both reference signals to identify the attended speaker. Section 6.5.2 describes the training step to compute the spatio-temporal filter.

6.5.1 *Decoding step*

C-channel EEG recordings $r_c[i]$, with $c = 1 \ldots C$ and $i = 1 \ldots I$ the sub-sampled time index, are first segmented into trials (see Section 6.6.3.4 for more details). An

estimate of the attended speech envelope $\hat{e}_a[i]$ using a trained spatio-temporal filter, i.e.,

$$\hat{e}_a[i] = \mathbf{g}^T \mathbf{r}[i], \quad i = 1 \dots I \tag{6.21}$$

with

$$\mathbf{g} = \left[\mathbf{g}_1^T \, \mathbf{g}_2^T \dots \, \mathbf{g}_C^T\right]^T, \tag{6.22}$$

$$\mathbf{g}_c = \left[g_{c,0} \; g_{c,1} \dots \; g_{c,J-1}\right]^T, \tag{6.23}$$

$$\mathbf{r}[i] = \left[\mathbf{r}_1^T[i] \; \mathbf{r}_2^T[i] \; \dots \; \mathbf{r}_C^T[i]\right]^T, \tag{6.24}$$

$$\mathbf{r}_c[i] = \left[r_c[i+\Delta] \; r_c[i+1+\Delta] \; \dots \; r_c[i+J-1+\Delta]\right]^T, \tag{6.25}$$

where J denotes the number of filter coefficients per channel, Δ modeling the latency of the attentional effect in the EEG responses to acoustic stimuli, and $(.)^T$ denotes the transpose operation. Next, the correlation coefficients ρ_1 and ρ_2 corresponding to speaker 1 and speaker 2 are computed by correlating the envelopes of both reference signals with the estimated attended speech envelope $\hat{e}_a[i]$, i.e.,

$$\rho_1 = \rho\left(e_1[i], \hat{e}_a[i]\right), \; \rho_2 = \rho\left(e_2[i], \hat{e}_a[i]\right), \tag{6.26}$$

where $e_1[i]$ and $e_2[i]$ denote the envelopes of the reference signals z_1 and z_2, respectively. Based on these correlation coefficients, it is then decided that the listener attended to speaker 1 if $\rho_1 > \rho_2$ or attended to speaker 2 otherwise. The correlation difference, defined as the difference between the correlation coefficients of the attended and the unattended speaker, i.e., $\Delta\rho = \rho\left(e_a[i], \hat{e}_a[i]\right) - \rho\left(e_u[i], \hat{e}_a[i]\right)$, can be used as a confidence measure for the AAD decision.

6.5.2 *Training step*

During the training step, the attended speaker is assumed to be known and an attended speech signal is used as the training signal. The filter $\mathbf{g}$ in (6.21) is computed by minimizing the least-squares error between the reconstructed envelope $\hat{e}_a[i]$ and the attended speech envelope $e_a[i]$, regularized with the squared l_2-norm of the derivatives of the filter coefficients to avoid over-fitting [16, 129, 169], i.e.

$$\min_{\mathbf{g}} \sum_{i=1}^{I} \left(e_a[i] - \mathbf{g}^T \mathbf{r}[i]\right)^2 + \beta \mathbf{g}^T \mathbf{D} \mathbf{g}, \tag{6.27}$$

where $\mathbf{D}$ denotes the derivative matrix [18] and β denotes a regularization parameter. The filter minimizing the regularized least-squares cost function in (6.27) is equal to

$$\mathbf{g} = \left(\mathbf{Q} + \beta \mathbf{D}\right)^{-1} \mathbf{q}, \tag{6.28}$$

with the correlation matrix $\mathbf{Q}$ and the cross-correlation vector $\mathbf{q}$ given by

$$\mathbf{Q} = \frac{1}{I} \sum_{i=1}^{I} \left(\mathbf{r}[i] \, \mathbf{r}^T[i]\right), \quad \mathbf{q} = \frac{1}{I} \sum_{i=1}^{I} \left(\mathbf{r}[i] \, e_a[i]\right). \tag{6.29}$$

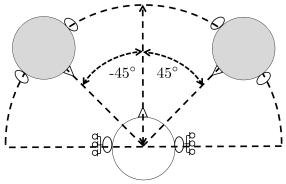


Fig. 6.2: Acoustic simulation setup. Two competing speakers were located at $\theta_1 = -45°$ and $\theta_2 = 45°$ with respect to the listener with two hearing aids, each equipped with 3 microphones.

In the simulations, we will consider the following training signals for computing the filter $\mathbf{g}$:

- anechoic training signal, i.e., the anechoic speech component of the attended speaker in the reference microphone signal at the side of the attended speaker.
- masked anechoic training signal using oracle binary masks of the attended speaker.

6.6 Experimental setup

In this section, we describe the acoustic simulation setup, the EEG measurement setup, the algorithm implementation details, and the used performance measures for AAD.

6.6.1 *Acoustic simulation setup*

Two highly overlapping German audio stories, uttered by two different male speakers, were used as the clean speech signals $s_1[n]$ and $s_2[n]$ (sampling frequency 16 kHz). The binaural hearing aid microphone signals $y_{L,m}[n]$ and $y_{R,m}[n]$ were generated by convolving the clean speech signals with non-individualized binaural impulse responses (anechoic or reverberant) from [110] and adding diffuse babble noise. The diffuse babble noise was simulated according to [192] using babble speech recordings and assuming a cylindrically isotropic noise field, i.e., using the anechoic ATFs from [110] with a resolution of 5°. The hearing aid setup in [110] consisted of two hearing aids, each equipped with $M = 3$ microphones, mounted on a dummy head. The left and the right competing speaker were located at $\theta_1 = -45°$ and $\theta_2 = 45°$ (see Fig. 6.2). In total, five acoustic conditions were considered: three anechoic conditions with binaural input SNRs ∞ dB, 4.0 dB and 9.0 dB, and two reverberant conditions (reverberation time $T_{60} \approx 0.5$ s) with binaural input SNRs 4.0 dB and 9.0 dB. The binaural input SNR (BSNR) is defined as the energy ratio between the speech components of speaker 1 and 2 in the reference microphone

signals and the background noise components in the reference microphone signals, i.e.,

$$\mathrm{BSNR}_{in} = 10\log_{10}\frac{E\left\{|x_{1,L,1}|^2\right\} + E\left\{|x_{2,L,1}|^2\right\} + E\left\{|x_{1,R,1}|^2\right\} + E\left\{|x_{2,R,1}|^2\right\}}{E\left\{|v_{L,1}|^2\right\} + E\left\{|v_{R,1}|^2\right\}}. \tag{6.30}$$

6.6.2 *EEG measurement*

Eighteen normal-hearing German-speaking participants took part in this study (see [26]). Before performing the experiment, the participants reported no, or very limited, knowledge of the audio stories. As acoustic stimuli, the reference microphone signals of the left and the right hearing aid were presented to the participants via insert earphones (E-A-RTONE 3A). Among all participants, 8 participants were instructed to attend to the left speaker, while 10 participants were instructed to attend to the right speaker. Two participants were excluded from the analysis, one participant due to a technical hardware issue since batteries of EEG amplifier suddenly went out of charge during EEG recording and the other participant due to poor attentional performance.

For all acoustic conditions, the EEG responses $r_c[i]$ were recorded using $C = 64$ channels (Easycap GmbH) at a sampling frequency of 500 Hz, and referenced to the nose electrode (see [26] for more details). Similarly as in [16,26], the EEG responses were re-referenced offline to a common average reference, band-pass filtered between 2 Hz and 8 Hz using a third-order Butterworth band-pass filter, and subsequently downsampled to 64 Hz.

6.6.3 *Algorithm implementation details*

6.6.3.1 *DSPP and DOA estimation (Section 6.3.1)*

For the DSPP and DOA estimation algorithm, the SVM classifiers were trained using simulated noisy speech signals, generated by convolving clean speech signals from the TIMIT database [223] with anechoic binaural impulse responses from [110] and adding diffuse speech-shaped noise at SNRs of -20 dB to 20 dB in steps of 10 dB. The GCC-PHAT features were calculated using a frame length of 10 ms with an overlap of 5 ms. The DSPPs $p_\theta[n]$ in (6.6) were recursively smoothed using a corresponding time constant of 1 s.

6.6.3.2 *Binary mask estimation (Section 6.4.1)*

The speech energies $\gamma_1[n]$ and $\gamma_2[n]$ were calculated using a frame length of 10 ms with an overlap of 5 ms. The energy thresholds $\alpha_{1,\gamma}$ and $\alpha_{2,\gamma}$ were determined as the median value of the speech energies $\gamma_1[n]$ and $\gamma_2[n]$, respectively. The DSPP

thresholds $\alpha_{1,p}$ and $\alpha_{2,p}$ were determined as the median of the DSPPs $\hat{p}_1[n]$ and $\hat{p}_2[n]$, respectively[1].

6.6.3.3 *MVDR beamforming (Section 6.3.2)*

The binaural MVDR beamformers were implemented using a weighted overlap-add (WOLA) framework with an STFT frame length of 512 samples and an overlap of 50%. The noise covariance matrix $\mathbf{\Phi}_{\mathbf{v}}$ was calculated assuming a cylindrically isotropic noise field, i.e., by spatially averaging the auto- and cross-correlations of the anechoic ATFs from [110] with a resolution of 5°. As proposed in [207], the anechoic RTF vector $\mathbf{a}(\theta)$ for angle θ was calculated as the normalized principal eigenvector of the (oracle) covariance matrix for angle θ, constructed using white noise convolved with the anechoic binaural impulse responses from [110] for angle θ. The anechoic RTF vector is normalized with respect to the left microphone when $\theta \leq 0°$ and with respect to the right microphone when $\theta > 0°$.

6.6.3.4 *AAD training and decoding (Section 6.5)*

The EEG recordings of the anechoic acoustic condition without background noise (SNR $= \infty$ dB) were considered for filter training, while the EEG recordings of the remaining four acoustic conditions were considered as unseen conditions for decoding. The acoustic conditions for decoding were grouped together based on reverberation time, resulting in two experimental analysis conditions, i.e., anechoic and reverberant. The EEG recordings corresponding to filter training were split into 80 trials, each of 30 seconds length. The EEG recordings corresponding to each experimental analysis condition were split into 40 trials, each of 30 seconds length. The attended speech envelope $e_a[i]$ used for filter training as well as the envelopes $e_1[i]$ and $e_2[i]$ of the reference signals used for decoding were obtained using a Hilbert transform [129], followed by low-pass filtering at 8 Hz and downsampling to 64 Hz. Each participant's own data were used for filter training and evaluation. Similarly as in [29], the parameters of the spatio-temporal filter $\mathbf{g}$ in (6.21) were set to $J = 8$ and $\Delta = 8$ (corresponding to 125 ms).

6.6.4 *AAD performance*

For each experimental condition, the following reference signals $z_1[n]$ and $z_2[n]$ were considered for decoding:

- non-masked reference signals, i.e., the oracle reference signals (ORACLE), the microphone signals (MIC), the MVDR ouput signals (MVDR).
- masked reference signals, i.e., the masked oracle reference signals using either ideal binary masks (IM-ORACLE) or estimated binary masks (EM-ORACLE),

1 Several criteria (mean, 25th percentile, median, 75th percentile, 90th percentile) were considered for the energy and DSPP thresholds. Among the considered criteria, the median yielded the largest decoding performance for the considered scenarios.

the masked microphone signals using either the ideal binary masks (IM-MIC) or the estimated binary masks (EM-MIC), and the masked MVDR output signals using either the ideal binary masks (IM-MVDR) or the estimated binary masks (EM-MVDR).

- binary mask reference signals, i.e., the ideal binary masks (IM) or the estimated binary masks (EM).

To evaluate the AAD performance with the non-masked reference signals, we used the spatio-temporal filter $\mathbf{g}$ trained with the anechoic training signal. To evaluate the AAD performance with the masked reference signals or with the binary mask reference signals, we used the spatio-temporal filter $\mathbf{g}$ trained with the masked anechoic training signal.

The AAD performance was computed by averaging the percentage of correctly decoded trials over all considered trails and all participants. In addition, as a confidence measure for AAD the correlation difference $\Delta\rho$ was averaged across all considered trails and participants. To statistically compare the decoding performance and the correlation difference, we performed a paired Wilcoxon signed rank test at the 5% significance level.

To analyze the performance for the correlation difference, we also considered the average signal-to-interference-plus-noise (SINR) in the reference signals, i.e.,

$$\mathrm{SINR} = \frac{\mathrm{SINR}_1 + \mathrm{SINR}_2}{2}, \tag{6.31}$$

with

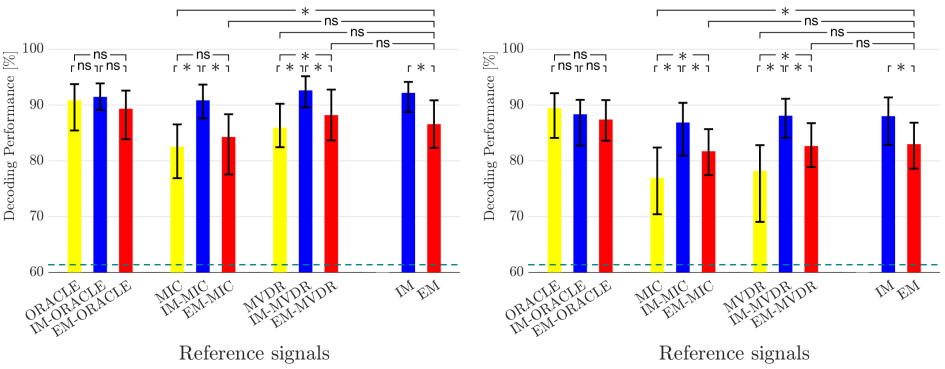
$$\mathrm{SINR}_1 = 10\log_{10}\frac{E\left\{|z_{1,1}|^2\right\}}{E\left\{|z_{1,2}|^2 + |z_{1,v}|^2\right\}}, \tag{6.32}$$

$$\mathrm{SINR}_2 = 10\log_{10}\frac{E\left\{|z_{2,2}|^2\right\}}{E\left\{|z_{2,1}|^2 + |z_{2,v}|^2\right\}}, \tag{6.33}$$

where $z_{1,1}$, $z_{1,2}$ and $z_{1,v}$ denote the speech component of speaker 1, the speech component of speaker 2 and the background noise component in the reference signal z_1 corresponding to speaker 1, respectively, and $z_{2,2}$, $z_{2,1}$ and $z_{2,v}$ denote the speech component of speaker 2, the speech component of speaker 1 and the background noise component in the reference signal z_2 corresponding to speaker 2, respectively.

6.7 Results and discussion

In this section, we evaluate the performance of the proposed reference signal generation algorithms using the experimental setup discussed in the previous section. In Section 6.7.1, we evaluate the decoding performance of the proposed algorithms. In Section 6.7.2, we analyze the performance of the proposed algorithms based on the correlation difference and SINR, where we also investigate the impact of DSPP estimation errors.



(a) anechoic condition (b) reverberant condition

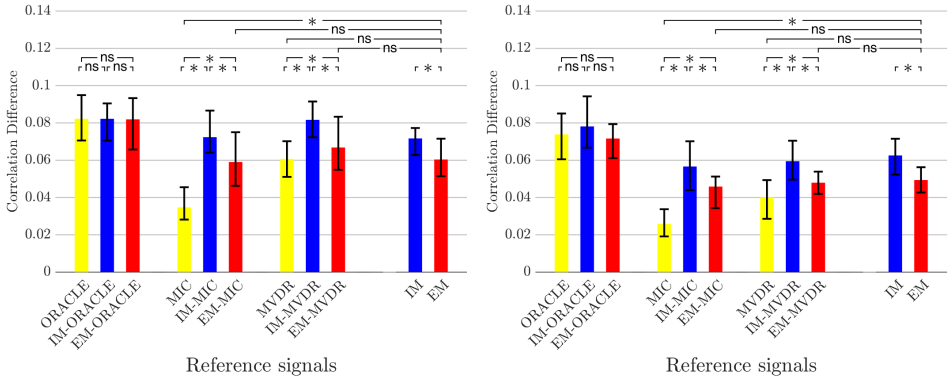
Fig. 6.3: Average decoding performance for (a) the anechoic condition and (b) the reverberant condition when using the oracle signals, the microphone reference signals, the MVDR reference signals, the masked reference signals either using ideal binary masks or estimated binary masks as reference signal for decoding. The red dashed-line represents the upper boundary of the confidence interval corresponding to chance level based on a binomial test at the 5% significance level, and $*$ indicates a significant difference ($p < 0.05$) based on the paired Wilcoxon signed rank test.

6.7.1 *Auditory attention decoding performance*

For the anechoic and the reverberant condition, Fig. 6.3 depicts the average decoding performance when using the non-masked reference signals, the masked reference signals and the binary mask reference signals.

When masking the oracle reference signals with ideal binary masks (IM–ORACLE), there is no significant difference in decoding performance compared to using the non-masked oracle reference signals (ORACLE). In addition, there is no significant difference in decoding performance between when masking the oracle reference signals with ideal binary masks (IM–ORACLE) and estimated binary masks (EM–ORACLE). These results show that discarding low speech energy intervals from the oracle reference signals (using either ideal or estimated binary masks) has hardly any impact on the decoding performance, indicating that the reference signals comprising moderate-to-high speech energy intervals of the oracle reference signals can be reliably used instead of the (complete) oracle reference signals for decoding auditory attention.

When masking either the microphone reference signals or the MVDR reference signals with ideal binary masks (IM–MIC or IM–MVDR), for both acoustic conditions the decoding performance significantly increases ($> 90.7\%$ for the anechoic condition and $> 86.8\%$ for the reverberant condition) compared to using the non-masked reference signals ($> 82.4\%$ for the anechoic condition and $> 76.8\%$ for the reverberant condition). When masking either the microphone reference signals or the



(a) anechoic condition (b) reverberant condition

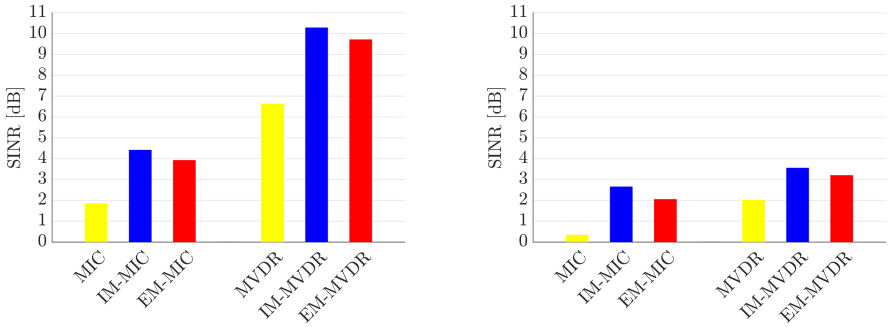
Fig. 6.4: Average correlation difference for (a) the anechoic condition and (b) the reverberant condition when using the non-masked reference signals, the masked reference signals and the binary mask reference signals, using ideal and estimated binary masks. The error bars represent the bootstrap confidence interval at the 5% significance level and $*$ indicates a significant difference ($p < 0.05$) based on the paired Wilcoxon signed rank test.

MVDR reference signals with estimated binary masks (EM–MIC or EM–MVDR) instead of ideal binary masks (IM–MIC, IM–MVDR), the decoding performance significantly decreases ($>$ 84.1% for the anechoic condition and $>$ 81.6% for the reverberant condition) but is still larger than the decoding performance using the non-masked reference signals especially for the reverberant condition.

When using the binary mask reference signals (IM or EM), it can be observed for both acoustic conditions that a large decoding performance is obtained. The decoding performance obtained by estimated binary masks is significantly lower (86.5% for the anechoic condition and 83.9% for the reverberant condition) compared to using ideal binary masks (92.1% for the anechoic condition and 87.9% for the reverberant condition). Nevertheless, the decoding performance obtained by estimated binary masks is comparable to masking the microphone reference signals or the MVDR reference signals with estimated binary masks. These results show that estimated binary masks can be reliably used instead of the masked microphone signals or the masked MVDR output signals as reference signals for decoding auditory attention.

6.7.2 *Correlation difference and SINR*

The decoding performance results in Fig. 6.3 can be further analyzed based on the correlation difference and SINR in Fig. 6.4 and Fig. 6.5. For both acoustic conditions, it can be observed that when masking the oracle reference signals with ideal binary masks (IM–ORACLE) there is no significant difference in correlation difference compared to using the non-masked oracle reference signals (ORACLE),



(a) anechoic condition (b) reverberant condition

Fig. 6.5: Average SINR of the non-masked and masked microphone and MVDR reference signals for (a) the anechoic condition and (b) the reverberant condition, using ideal and estimated binary masks.

consistent with the decoding performance results in Fig. 6.3. In addition, there is no significant difference in correlation difference between masking the oracle reference signals with ideal binary masks (IM–ORACLE) and estimated binary masks (EM–ORACLE). The comparable correlation difference obtained by the oracle reference signals and masking with ideal or estimated binary masks may be explained by considering the auditory neural responses to landmarks of high speech energy such as speech envelope peaks and syllable onsets. Several studies have shown that the neural responses are correlated with high-energy landmarks [224–226], during which the speech energy and the speech presence probability are high. Discarding intervals with low speech energy either using ideal or estimated binary masks results in signals comprising intervals with moderate-to-high speech energy, including the high-energy speech landmarks.

When masking either the microphone reference signals or the MVDR reference signals with ideal binary masks (IM–MIC or IM–MVDR), for both acoustic conditions a significantly larger correlation difference is obtained compared to using the non-masked reference signals (MIC or MVDR). When masking with estimated binary masks (EM–MIC or EM–MVDR) instead of ideal binary masks (IM–MIC, IM–MVDR), the correlation difference significantly decreases. Nevertheless, the correlation difference obtained by estimated binary masks is still larger than using the non-masked reference signals, as reflected by larger corresponding decoding performances in Fig. 6.3. The correlation differences obtained by masking with ideal binary masks and estimated binary masks can be explained by considering the SINR resulted in Fig. 6.5. It can be observed that when masking either the microphone reference signals or the MVDR reference signals with ideal binary masks or estimated binary masks, a larger SINR is obtained (about 3.9 - 10.3 dB for the anechoic condition and 2.0 - 3.5 dB for the reverberant condition) compared to using the non-masked reference signals (1.8 dB for the anechoic condition and 0.3 dB for the reverberant condition). Moreover, it can be observed that the SINR obtained

by masking with ideal binary masks (about 4.4 - 10.3 dB for the anechoic condition and 2.6 - 3.5 dB for the reverberant condition) is larger than using estimated binary masks (about 3.9 - 9.7 dB for the anechoic condition and 2.0 - 3.2 dB for the reverberant condition). These results show that the correlation difference and the SINR become sensitive to DSPP estimation errors when masking either the microphone reference signals or the MVDR reference signals.

When using the binary mask reference signals, it can be observed for both acoustic conditions that the correlation difference is significantly larger than zero, which is reflected by the significant decoding performances in Fig. 6.3. The fact that a significant correlation difference is obtained using the binary mask reference signals may be explained by considering a neural representation of high-energy landmarks. In [226, 227] it has been shown that the human brain, especially the superior temporal gyrus, encodes high-energy speech landmarks rather than a complete speech envelope. Since the envelopes of the binary mask reference signals comprise landmarks of moderate-to-high speech energy, these envelopes are expected to be correlated with auditory neural responses.

6.8 Conclusion

In this paper, we proposed reference signal generation algorithms based on binary masking, which discard low-energy intervals susceptible to interfering speech and background noise. The proposed algorithms determine the binary masks from the DSPP and the DOAs of both speakers, which are estimated from the microphone signals. The reference signals for decoding are then generated by either masking the microphone signals or the MVDR output signals. The reference signals comprising moderate-to-high speech energy intervals of the oracle reference signals generated by masking were also considered. Moreover, the binary masks were considered themselves as reference signals for decoding. The performance of the proposed reference signal generation algorithms was evaluated for a two-speaker scenario in diffuse babble noise. The experimental results showed that the reference signals comprising moderate-to-high speech energy intervals of the oracle reference signals can be reliably used instead of the (complete) oracle reference signals for decoding auditory attention. This suggests that for generating appropriate reference signals, e.g., from the microphone signals, it is more important to preserve the moderate-to-high speech energy intervals than low speech energy intervals. Moreover, the experimental results showed that the proposed algorithms masking either the microphone signals or the MVDR output signals significantly improve the decoding performance compared to when using the (non-masked) microphone and MVDR output signals especially in the reverberant condition. Quite remarkably, the results showed that AAD using the binary masks as reference signals yields a decoding performance that is comparable to using the masked MVDR output signals as reference signals.

7

IMPROVING AUDITORY ATTENTION DECODING PERFORMANCE OF LINEAR AND NON-LINEAR METHODS USING STATE-SPACE MODEL

Identifying the target speaker in hearing aid applications is crucial to improve speech understanding. Recent advances in electroencephalography (EEG) have shown that it is possible to identify the target speaker from single-trial EEG recordings using auditory attention decoding (AAD) methods. AAD methods reconstruct the attended speech envelope from EEG recordings, based on a linear least-squares cost function or non-linear neural networks, and then directly compare the reconstructed envelope with the speech envelopes of speakers to identify the attended speaker using Pearson correlation coefficients. Since these correlation coefficients are highly fluctuating, for a reliable decoding a large correlation window is used, which causes a large processing delay. In this paper, we investigate a state-space model using correlation coefficients obtained with a small correlation window to improve the decoding performance of the linear and the non-linear AAD methods. The experimental results show that the state-space model significantly improves the decoding performance.

7.1 Introduction

Multi-microphone speech enhancement algorithms in currently available hearing aid devices are able to perform source separation and reduce background noise to improve speech intelligibility. However, the performance of these algorithms in improving speech intelligibility typically depends on correctly identifying the target speaker to be enhanced. In hearing aid applications, the target speaker is typically identified using assumptions such as the target speaker being located in front of the listener or being the loudest speaker. However, since in real-world conditions these assumptions may often be violated, e.g., when the auditory attention of the listener is misaligned with the assumptions, the performance of speech enhancement methods decreases and results in a substantially reduced benefit from hearing aids. Recent advances in electroencephalography (EEG) have shown that it is possible to identify the target speaker from single-trial EEG recordings of a listener by decoding

the auditory attention [27,28,31]. Several auditory attention decoding (AAD) methods have been proposed to identify the target speaker, based on, e.g., a linear least-squares cost function [16, 22, 24, 128] and non-linear neural networks [23, 130]. The linear least-squares-based AAD method proposed in [16] is able to exploit the linear neural process of attention along the auditory pathway to identify the attended speaker. The non-linear neural-network-based AAD method proposed in [23] is able to exploit the non-linear neural process of attention in addition to the linear neural process. To identify the attended speaker, these methods aim at reconstructing the attended speech envelope from the EEG recordings using a trained spatio-temporal estimator. In the training step, a spatio-temporal envelope estimator is trained by either minimizing the least-squares error or maximizing the correlation cost function between the attended speech envelope and the reconstructed envelope. In the decoding step, the attended speech envelope is reconstructed using the trained envelope estimator and then directly compared with the speech envelopes of two speakers using Pearson correlation coefficients to identify the attended speaker. Since these correlation coefficients are highly fluctuating, for a reliable decoding a large correlation window on the order of 30 seconds is typically used, which causes a large processing delay and hence limits the feasibility of AAD for hearing aid applications. In [24], it has been proposed to use coefficients of least-squares-based envelope estimators, obtained separately for reconstructing the attended and the unattended speech envelope. Using coefficients of estimators, a state-space model then identifies the attended speaker. In this paper, we investigate a state-space model using correlation coefficients obtained with a small correlation window to improve the decoding performance of the (linear) least-squares-based AAD method and the (non-linear) neural-network-based AAD method. The correlation coefficients are generated using either the least-squares-based AAD method or the neural-network-based AAD method. The state-space model then translates the generated correlation coefficients into smooth estimates of the attention state, based on which the attended speaker is identified.

For an acoustic scenario with two competing speakers and diffuse noise at different SNRs and reverberation times, 64-channel EEG responses with 18 participants were recorded. The experimental results show for correlation coefficients obtained with a 5-second correlation window that the least-squares-based AAD method and the neural-network-based AAD method yield a low decoding performance. However, when using the state-space model with the least-squares-based AAD method, the decoding performance significantly improves.

7.2 Auditory attention decoding

This section presents the auditory attention decoding using a state-space model, which employs correlation coefficients generated either by the least-squares-based AAD method and the neural-network-based AAD method. In Section 7.2.1 the acoustic scenario and the notation are defined. Section 7.2.2 describes the state-space model. Section 7.2.3 and Section 7.2.4 describe the least-squares-based AAD method and the neural-network-based AAD method.

7.2.1 *Configuration and notation*

We consider an acoustic scenario comprising two competing speakers and background noise in a reverberant environment, where the ongoing EEG responses of a listener to these acoustic stimuli are recorded (See Fig. 7.1). The clean speech signal of speaker 1 is denoted as $s_1[n]$, with n the discrete time index, while the clean speech signal of speaker 2 is denoted as $s_2[n]$. The envelopes of the clean speech signals of speaker 1 and 2 are denoted as $e_1[k]$ and $e_2[k]$, with k the sub-sampled time index, respectively.

The reconstructed attended speech envelope from C-channel EEG recordings $r_c[k]$, with $c = 1 \dots C$, using a trained spatio-temporal envelope estimator F is given by

$$\hat{e}_a[k] = F(\mathbf{r}[k]), \tag{7.1}$$

with

$$\mathbf{r}[k] = \left[\mathbf{r}_1^T[k]\ \mathbf{r}_2^T[k]\ \dots\ \mathbf{r}_C^T[k]\right]^T, \tag{7.2}$$

$$\mathbf{r}_c[k] = \left[r_c[k]\ r_c[k+1]\ \dots\ r_c[k+\Delta]\right]^T, \tag{7.3}$$

where Δ denotes the latency considered for modeling the attentional effect in the EEG responses to acoustic stimuli.

The Pearson correlation coefficients between the reconstructed attended envelope $\hat{e}_a[k]$ and the envelope of two speakers are given by

$$\rho_{1,k} = \rho\left(\mathbf{e}_1[k],\ \hat{\mathbf{e}}_a[k]\right), \quad \rho_{2,k} = \rho\left(\mathbf{e}_2[k],\ \hat{\mathbf{e}}_a[k]\right), \tag{7.4}$$

where $\hat{\mathbf{e}}_a[k]$ denotes the stacked vector of the reconstructed attended envelope corresponding to a correlation window of length K_{COR}, i.e.,

$$\hat{\mathbf{e}}_a[k] = \left[\hat{e}_a[(k-1)K_{\mathrm{COR}}+1]\ \hat{e}_a[(k-1)K_{\mathrm{COR}}+2]\ \dots\ \hat{e}_a[kK_{\mathrm{COR}}]\right]^T, \tag{7.5}$$

and $\mathbf{e}_1[k]$ and $\mathbf{e}_2[k]$ are defined similarly as in (7.5). Please note that in this paper we assume that the clean speech signal of speakers are available for obtaining the envelopes of speakers $e_1[k]$ and $e_2[k]$. However, since in practice only microphone signals containing a mixture of speakers and ambient noise are available, the clean speech signal of speakers needs to be appropriately estimated from microphone signals, e.g., by using the noise reduction and source separation algorithms proposed in [27,31,170,212].

7.2.2 *AAD using state-space model*

Suppose the attended envelope is reconstructed using a trained (linear or nonlinear) spatio-temporal estimator and the correlation coefficients of speakers are obtained. We aim at estimating the probability of attending to speaker 1 or 2 based on a state-space model using the past and the subsequent correlation coefficients (see

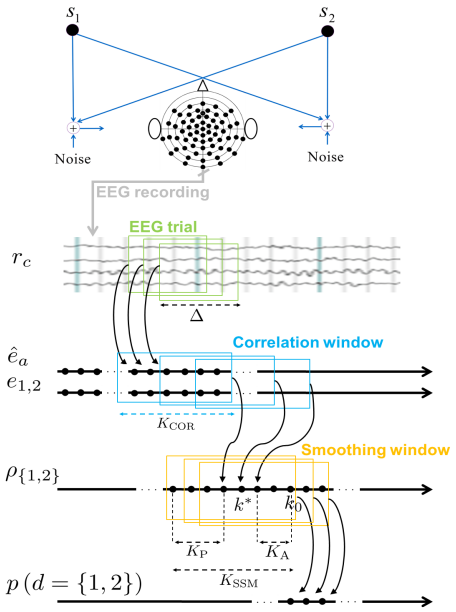


Fig. 7.1: Illustration of the process flow of AAD using state-space model.

Fig. 7.1). Let the attention state of the listener when attending to either speaker 1 or 2 be defined as a binary random variable, i.e.,

$$\begin{cases} d_k = 1, & \text{when attending to speaker 1} \\ d_k = 2, & \text{when attending to speaker 2} \end{cases}, \tag{7.6}$$

which follows a Bernoulli process. The probability of attending to speakers based on the state-space model is obtained as [24, 197]

$$p(d_k = 1) = 1 - p(d_k = 2) = \frac{1}{1 + e^{-(z_k)}}, \tag{7.7}$$

with

$$z_k = c_0 z_{k-1} + w_k, \tag{7.8}$$

$$w_k \sim \mathcal{N}(0, \eta_k), \tag{7.9}$$

$$\eta_k \sim \text{Inverse-Gamma}(a_0, b_0), \tag{7.10}$$

, c_0 denoting the hyperparameter ensuring the stability of z_k, and a_0 and b_0 denoting the hyperparameters used to control the smoothing degree of the state-space model by tuning the variations of z_k and $p(d_k = \{1, 2\})$. The autoregressive model in (7.8) implies that the (attention state) parameter z_k at instance k is predicted from z_{k-1} at the instance $k-1$ with some uncertainty, which is modeled by the noise process $w(k)$. Please note that when z_k varies from $-\infty$ to ∞, $p(d_k = 1)$ monotonically varies from 0 to 1. To relate the correlation coefficients of speakers to the attention

state, the probability of the absolute values of correlation coefficients given attending to either speaker 1 or 2 is modeled using a Log-Normal distribution, i.e.,

$$p\left(|\rho_{l,k}| \mid d_k = l\right) \sim \text{Log-Normal}\left(\boldsymbol{\alpha}_a\right), \quad l = 1, 2 \tag{7.11}$$

with $\boldsymbol{\alpha}_a$ denoting the parameter set of the attended Log-Normal distribution. The probability of the correlation coefficients given ignoring either speaker 1 or 2 is modeled as

$$p\left(|\rho_{l,k}| \mid d_k \neq l\right) \sim \text{Log-Normal}\left(\boldsymbol{\alpha}_u\right), \quad l = 1, 2 \tag{7.12}$$

with $\boldsymbol{\alpha}_u$ denoting the parameter set of the unattended Log-Normal distribution.

Let's suppose we are at the instance $k = k_0$ (see Fig. 7.1) and aim to estimate the probability of attending to speakers $p\left(d_k = \{1, 2\}\right)$ at the instance $k = k^*$ using the correlation coefficients obtained within a sliding smoothing window of length $K_{\text{SSM}} = K_P + K_A + 1$, with K_P and K_A denoting the parameters determining the number of the correlation coefficient prior to and after the instance k^*, respectively. The parameters of the state-space model corresponding to the smoothing window are hence given as $\Omega = \{z_{k_0-K_{\text{SSM}}+1:k_0}, \eta_{k_0-K_{\text{SSM}}+1:k_0}, \boldsymbol{\alpha}_a, \boldsymbol{\alpha}_u\}$. Theses parameters including z_{k^*} are estimated from the correlation coefficients $\rho_{1,k_0-K_{\text{SSM}}+1:k_0}$ and $\rho_{2,k_0-K_{\text{SSM}}+1:k_0}$ obtained within the smoothing window using the Expectation Maximization (EM) estimation algorithm proposed in [24, 197]. Based on the estimated attention state parameter z_{k^*}, the probability of attending to speakers $p\left(d_{k^*} = \{1, 2\}\right)$ are obtained. It is then decided that the listener attended to speaker 1 if $p\left(d_{k^*} = 1\right) > p\left(d_{k^*} = 2\right)$ or attended to speaker 2 otherwise. Please note that the estimated parameters Ω are also used for the initialization of parameters in the next smoothing window.

In the simulations (see Section 7.3), we will consider to use the state-space model with correlation coefficients generated either by the least-squares-based AAD method (see Section 7.2.3) or the neural-network-based AAD method (see Section 7.2.4).

7.2.3 *Least-squares-based AAD*

The least-squares-based AAD method proposed in [16] aims at estimating the attended speech envelope from the EEG recordings using a trained linear spatio-temporal estimator. In the training step, the attended speaker is assumed to be known and an attended speech signal is used to train a linear estimator by minimizing the least-squares error between the attended speech envelope $e_a\left[k\right]$ and the reconstructed envelope $\hat{e}_a\left[k\right]$, i.e.,

$$\min_{\mathbf{g}} \frac{1}{K} \sum_{k=1}^{K} \left(e_a\left[k\right] - \hat{e}_a\left[k\right]\right)^2 + \beta \mathbf{g}^T \mathbf{D} \mathbf{g}, \tag{7.13}$$

with $\hat{e}_a[k] = F(\mathbf{r}[k]) = \mathbf{g}^T\mathbf{r}[k]$, $\mathbf{D}$ denoting the derivative matrix [169] and β denoting a regularization parameter. The linear estimator minimizing the regularized least-squares cost function in (7.13) is equal to

$$\mathbf{g} = (\mathbf{Q} + \beta\mathbf{D})^{-1}\,\mathbf{q}, \tag{7.14}$$

with the correlation matrix $\mathbf{Q}$ and the cross-correlation vector $\mathbf{q}$ given by

$$\mathbf{Q} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{r}[k]\,\mathbf{r}^T[k]\right), \quad \mathbf{q} = \frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{r}[k]\,e_a[k]\right). \tag{7.15}$$

In the decoding step, the attended envelope $\hat{e}_a[k]$ is obtained using the (trained) linear estimator $\mathbf{g}$ in (7.14). Next, the correlation coefficients between the reconstructed attended envelope and the envelope of two speaker $\rho_{1,k}$ and $\rho_{2,k}$ are computed as in (7.4). Based on these correlation coefficients, it is then decided that the listener attended to speaker 1 if $\rho_{1,k} > \rho_{2,k}$ or attended to speaker 2 otherwise.

7.2.4 *Neural-network-based AAD*

The neural-network-based AAD method aims at estimating the attended speech envelope from the EEG recordings using a trained non-linear spatio-temporal estimator. Similarly as in [23, 130], we consider a network $\mathcal{H}$ consisting of a hidden convolutional layer with hyperbolic tangent activation functions and one output layer with linear activation functions. In the training step, the network is trained to maximize the correlation between the attended speech envelope and the reconstructed envelope by minimizing the correlation cost function [23], i.e.,

$$\min\ \frac{1}{K}\sum_{k=1}^{K}(1 - \rho\left(\mathbf{e}_a[k]\,,\ \hat{\mathbf{e}}_a[k]\right)), \tag{7.16}$$

A correlation cost function equal to 0 corresponds to the maximum correlation between the attended speech envelope and the reconstructed envelope, i.e., $\rho\left(\mathbf{e}_a[k]\,,\ \hat{\mathbf{e}}_a[k]\right) = 1$, while a correlation cost function equal to 1 corresponds to the minimum correlation. A correlation cost function larger than 1 corresponds to a negative correlation.

In the decoding step, the attended envelope is obtained using the (trained) network $\mathcal{H}$, i.e., $\hat{e}_a[k] = F(\mathbf{r}[k]) = \mathcal{H}(\mathbf{r}[k])$. Next, the correlation coefficients between the reconstructed attended envelope and the envelope of two speaker are computed $\rho_{1,k}$ and $\rho_{2,k}$ as in (7.4). Based on these correlation coefficients, it is then decided that the listener attended to speaker 1 if $\rho_{1,k} > \rho_{2,k}$ or attended to speaker 2 otherwise.

7.3 Experimental setup

7.3.1 *Acoustic stimuli and EEG measurement*

EEG responses were recorded for 18 native German-speaking participants. Two German audio stories, uttered by two different male speakers, were simultaneously presented to the participants using insert earphones. The presented stimuli at both ears were generated by convolving the clean speech signals, i.e., the audio stories, with (non-individualized) binaural impulse responses from [110], and adding diffuse noise, generated according to [192]. The left and the right speaker were simulated at $\theta_1 = -45^\circ$ and $\theta_2 = 45^\circ$. Eight different acoustic conditions were considered for the stimuli: one anechoic condition with no background noise, two reverberant conditions with a moderate and a large reverberation time (reverberation time $T_{60} =$ 0.5 s and 1 s), two anechoic conditions with binaural input SNRs = 9.0 dB and 4.0 dB, and three combinations of reverberation and noise. Among all participants, 8 participants were instructed to attend to the left speaker, while 10 participants were instructed to attend to the right speaker. Two participants were excluded from the analysis, one participant due to poor attentional performance and the other one due to a technical hardware problem. The EEG responses were recorded using $C = 64$ channels at a sampling frequency of 500 Hz, and referenced to the nose electrode. The EEG responses were re-referenced offline to a common average reference, band-pass filtered between 2 Hz and 8 Hz using a third-order Butterworth band-pass filter, and subsequently downsampled to 64 Hz. The envelopes of the speech signals were obtained using a Hilbert transform, followed by low-pass filtering at 8 Hz and downsampling to 64 Hz.

7.3.2 *AAD training and testing*

For AAD training and testing, the EEG recordings for the different acoustic conditions were grouped together based on acoustic similarity, resulting in four experimental analysis conditions, i.e., anechoic, reverberant, anechoic-noisy, and reverberant-noisy, each of length 20 minutes. To avoid using EEG recordings of the same experimental analysis condition for training and testing, the leave-one-condition-out approach was used, i.e., four combinations of three experimental analysis conditions without repetition were considered for training and the left condition for each combination was considered for testing. This resulted in four training conditions and four testing conditions.

For the least-squares-based AAD method, the latency parameter of the linear estimator in (7.3) was set to $\Delta = 20$ (corresponding to 312 ms), as found to be an appropriate choice for AAD [16, 26]. For training, the estimator in (7.14) and the regularization parameter β of the estimator in (7.13) was determined using a k-fold cross-validation approach with $k = 10$, each of length 6 minutes. For testing, the EEG recordings were segmented into trials of length 5 s with an overlap of 4.98 s (corresponding to one sample shift). The correlation coefficients were computed using a correlation window of length $K_{\mathrm{COR}} = 5$ s with an overlap of 4.5 s.

For the neural-network-based AAD method, the network $\mathcal{H}$ consisting of a convolutional hidden layer with a filter kernel size of 20 samples (corresponding to 312 ms) was used. For training, the network was trained using a k-fold cross-validation approach with $k = 10$, each of length 6 minutes. The network was trained with the Nadam optimizer [228] using a batch size of 3840 samples (corresponding to 60 seconds $\times$ 64 channels), a learning rate of 0.002, and 3000 iterations. To avoid over-fitting, the dropout technique from [229] was used with a ratio of 0.25, which corresponds to randomly setting 25% of the hidden units to 0. The network was implemented in Keras [230]. For testing, the correlation coefficients were obtained using the same correlation window setting as used for the least-squares-based AAD method.

For the state space model, the hyperparameters c_0 in (7.8) and a_0 and b_0 in (7.10) were set to $c_0 = 1$, $a_0 = 2.008$ and $b_0 = 0.2016$, similarly as in [24]. For testing, a sliding smoothing window of length $K_{\text{SSM}} = 3$ with $K_A = 1$, $K_P = 1$ was used. For each testing condition, the parameter set of the attended Log-Normal distribution $\boldsymbol{\alpha}_a$ in (7.11) was initialized by fitting over correlation coefficients of the (oracle) attended speaker obtained during the first 15 s and was then fixed. The parameter set of the unattended Log-Normal distribution $\boldsymbol{\alpha}_u$ in (7.12) was similarly initialized by fitting over correlation coefficients of the (oracle) unattended speaker. For testing, the parameters of the state-space model Ω corresponding to an smoothing window were estimated using the EM estimation algorithm with 20 iterations.

The quality of correlation coefficients generated by either the least-squares-based AAD method or the neural-network-based AAD method was evaluated in terms of the attended correlation and the unattended correlation. The attended correlation was computed using the Pearson correlation between the reconstructed envelopes and the envelopes of the attended speaker, i.e., $\rho_{a,k} = \rho\left(\mathbf{e}_a\left[k\right],\ \hat{\mathbf{e}}_a\left[k\right]\right)$. The unattended correlation was computed between the reconstructed envelopes and the envelopes of the unattended speaker, i.e. $\rho_{u,k} = \rho\left(\mathbf{e}_u\left[k\right],\ \hat{\mathbf{e}}_a\left[k\right]\right)$,

The decoding performance was evaluated for several AAD methods, i.e., the least-squares-based AAD method (LS), the neural-network-based AAD method (NN), the state-space model using with the least-squares-based AAD method (LS-SSM) and the state-space model with the neural-network-based AAD method (NN-SSM). The decoding performance for the least-squares-based and the neural-network-based AAD method was computed as the percentage of correctly decoded 5-second correlation windows. The decoding performance for the state-space model using either the least-squares-based or the neural-network-based AAD method was computed as the percentage of correctly decoded 5-second smoothing windows.

7.4 Results and discussion

For the least-squares-based AAD method and the neural-network-based AAD method, Fig. 7.2 depicts the attended correlation and the unattended correlation for different acoustic conditions. It can be observed for all acoustic conditions that the attended correlation obtained by the neural-network-based AAD method (NN) is larger than the least-squares-based AAD method (LS), showing that the neural-

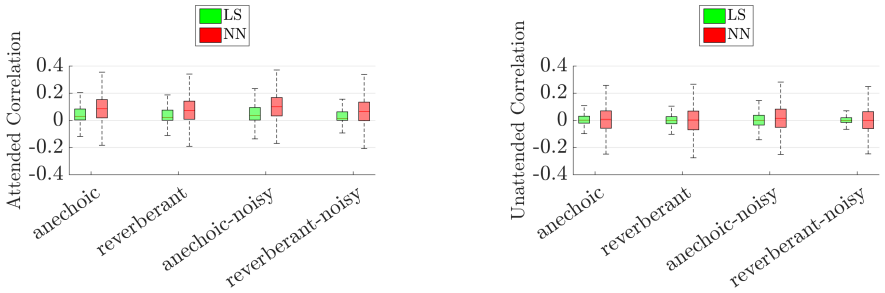


Fig. 7.2: Attended correlation and unattended correlation for different acoustic conditions when using the least-squares-based method and the neural-network-based AAD method.

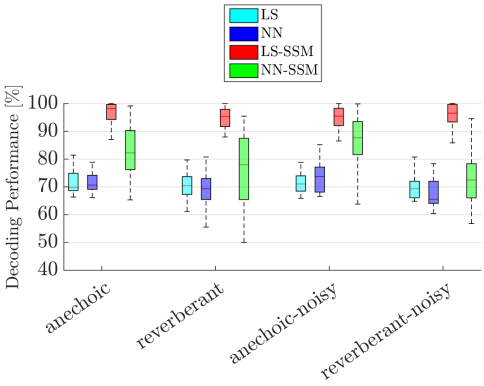


Fig. 7.3: Decoding performance for different acoustic conditions when using the least-squares-based method, the neural-network-based AAD method, the state-space model with the least-squares-based method and the state-space model with the neural-network-based AAD method.

network-based AAD method is able to reconstruct the attended speech envelope with a better accuracy. However, the attended correlation obtained by the neural-network-based AAD has a larger variability compared to the least-squares-based AAD method, which corresponds to attended correlation coefficients with a larger fluctuation. In addition, it can be observed that there is no significant difference in the unattended correlation obtained by the least-squares-based AAD method and the neural-network-based AAD method. However, the unattended correlation obtained by the neural-network-based AAD has a larger variability compared to the least-squares-based AAD method.

For all acoustic conditions, Fig. 7.3 depicts the decoding performance when using either the least-squares-based AAD method, the neural-network-based AAD method, the state-space model with the least-squares-based AAD method, or the state-space model with neural-network-based AAD method. It can be observed that when us-

ing either the least-squares-based AAD method or the neural-network-based AAD method, a relatively low decoding performance (with the median decoding performance $69\% - 73\%$) is obtained, mainly due to quite small (attended and unattended) correlations with a large variability (see Fig. 7.2) based on which decoding is performed by these methods. A statistical multiple comparison test (Kruskal-Wallis test followed by the posthoc Dunn and Sidak test [1]) revealed no significant difference ($p > 0.05$) in decoding performance when using the least-squares-based AAD method or the neural-network-based AAD method. When using the state-space model with either the least-squares-based or the neural-network-based AAD method (LS–SSM, LS–NN), the decoding performance increases. The increase is considerably larger for the least-squares-based AAD method (with the median decoding performance $> 94\%$) compared to the neural-network-based AAD method (with the median decoding performance $> 73\%$). The larger decoding performance can be explained by the fact that the correlations generated by the least-squares-based AAD method have a lower variability compared to the correlations generated by the neural-network-based AAD method, which leads to a smoother estimate of attention probabilities and a more stable decoding. The statistical multiple comparison test revealed that for most acoustic conditions (except anechoic–noisy) the decoding performance using the state-space model with the least-squares-based AAD method is significantly larger ($p < 0.05$) compared to using the least-squares-based AAD method, the neural-network-based AAD method, and the state-space model with the neural-network-based AAD method.

7.5 Conclusion

In this paper, we investigated a state-space model using correlation coefficients obtained with a 5-second correlation window to improve the decoding performance of the (linear) least-squares-based AAD method and the (non-linear) neural-network-based AAD method. The state-space model translates correlation coefficients, generated either by the least-squares-based or the neural-network-based AAD method, into smooth estimates of the attention state. The experimental results showed for all acoustic conditions that there is no significant difference in decoding performance between using the least-squares-based AAD method and the neural-network-based AAD method. However, when using the state-space model with the least-squares-based AAD method, for most acoustic conditions the decoding performance significantly improves.

8

CONCLUSION AND FURTHER RESEARCH

In this chapter, we provide a summary of the main contributions of the thesis in Chapters 2−7 and discuss possible extensions and directions for further research which could be envisaged as a follow-up to the work presented in this thesis.

8.1 Conclusion

Identifying the target speaker in hearing aid applications is an essential ingredient to improve speech intelligibility. Although several speech enhancement algorithms are available to reduce background noise or to perform source separation in multi-speaker scenarios, their performance depends on correctly identifying the target speaker to be enhanced. Recent advances in EEG have shown that it is possible to identify the target speaker which the listener is attending to using single-trial EEG-based AAD methods. However, in realistic acoustic environments the performance of AAD is influenced by undesired disturbances (interfering speakers, noise and reverberation), which have a negative effect on the reference signals used for decoding auditory attention. Additionally, it is important for real-world applications to close the loop by presenting on-line auditory feedback according to the AAD results.

The main objectives of this thesis were to analyze the performance of AAD for a two-speaker scenario in realistic noisy and reverberant acoustic conditions, to improve AAD methods and to develop and evaluate open-loop and closed-loop cognitive-driven speech enhancement systems for hearing aid applications using AAD. In Chapter 2, we analyzed the performance of a least-squares-based AAD method by investigating the impact of different acoustic conditions (anechoic, reverberant, noisy, and reverberant-noisy) for AAD filter training and for decoding as well as by investigating the influence of different acoustic disturbances on the reference signals used for AAD. In Chapters 3, 4 and 5, we proposed open-loop and closed-loop cognitive-driven speech enhancement systems to enhance the attended speaker while controlling the suppression of the interfering speaker. In Chapters 6 and 7, we proposed AAD methods based on either binary masking or a state-space model.

In **Chapter 2**, we investigated the performance of the least-squares-based AAD method for different acoustic conditions, both in the training step as well as in the decoding step (using correlation coefficients obtained with a 60-second correlation window). The experimental results showed that for all considered acoustic

conditions it is possible to decode auditory attention with a considerably large decoding performance ($> 90\%$[1]), even when the acoustic conditions for training and decoding are different. In addition, for most acoustic conditions there is no significant difference in decoding performance when using filters trained in all conditions or filters trained in a specific condition. This suggests that for an unseen realistic acoustic condition AAD can be performed using filters trained in, e.g., a laboratory acoustic condition. Furthermore, we investigated the influence of the head filtering effect and of acoustic components (reverberation, background noise and interfering speaker) on the decoding performance. The experimental results showed that for all considered acoustic conditions the head filtering effect has no significant impact on the decoding performance. Moreover, when using speech signals affected by either reverberation or background noise as reference signals, a comparable decoding performance is obtained as when using clean speech signals as reference signals. On the contrary, when using speech signals affected by the interfering speaker as reference signals, the decoding performance significantly decreases ($> 87\%$). This suggests that for generating appropriate reference signals, e.g., using acoustic signal pre-processing algorithms, it is more important to reduce the interfering speaker than to reduce background noise or reverberation. Furthermore, when using the (unprocessed) noisy and reverberant microphone signals as reference signals for decoding, a relatively large decoding performance (87%) can still be obtained. This implies that decoding is feasible for the considered scenario even based on the unprocessed noisy and reverberant signals. Moreover, we explored the potential of using the attended speech signal affected by different acoustic components as training signal for computing the filter. When using attended speech signals affected by either reverberation or by background noise as training signal, a comparable decoding performance is obtained as when using the clean attended speech signal as training signal ($> 92\%$). However, when using attended speech signals affected by the interfering speaker as training signal, the decoding performance significantly decreases ($> 82\%$). Nevertheless, even when using the attended speech signals affected by reverberation, background noise and an interfering speaker as training signal, it is still feasible to achieve a large decoding performance (82%).

Aiming at enhancing the attended speaker and controlling the suppression of the unattended speaker while preserving the spatial impression of the acoustic scene, in **Chapter 3** we proposed an open-loop cognitive-driven binaural speech enhancement system. The system cognitively steers a binaural MVDR or LCMV beamformer based on AAD and estimated (anechoic or reverberant) RTFs. Based on the estimated RTFs of both speakers, two MVDR or LCMV beamformers generate reference signals for auditory attention decoding. Using the envelopes of these reference signals, the EEG recordings and a 30-second correlation window, in the AAD step the DOAs of the attended and the unattended speaker are identified to steer the binaural MVDR or LCMV beamformer. The experimental results showed that for a two-speaker scenario in diffuse babble noise the proposed system using

1 Please note that the decoding performance reported in Chapter 2 was obtained using spatio-temporal AAD filters with parameters optimized per participant.

anechoic RTFs significantly improves the binaural SINR for the anechoic condition (7.4–8.7 dB) as well as for the reverberant condition (3.2–3.6 dB) compared to a fixed forward-steered binaural MVDR beamformer (0.3 dB for the anechoic condition and 0.5 dB for the reverberant condition). In particular, the cognitive-driven binaural LCMV beamformer (with the interference suppression factor $\delta = 0.1$) is able to both improve the binaural SINR as well as preserve the binaural cues of both the attended and the unattended speaker. Moreover, the results showed that the proposed system using estimated anechoic RTFs yields a larger binaural SINR improvement and AAD performance for the reverberant condition compared to using estimated reverberant RTFs. In addition, the results for both MVDR and LCMV beamformers showed that when using estimated anechoic RTFs the decoding performance is larger than when using the unprocessed microphone signals ($> 77\%$[2]). The decoding performance for the LCMV beamformers ($> 82\%$) is larger than for the MVDR beamformers ($> 78\%$), which can be explained by the larger SINR improvement of the LCMV beamformers and especially the larger interference suppression compared to the MVDR beamformers. This is in accordance with the experimental results in Chapter 2, where it was shown that jointly suppressing interference and background noise is of great importance for reference signal generation. Moreover, the results showed that for the considered experimental setup the AAD performance and the binaural SINR improvement of the proposed system are sensitive to RTF estimation errors and AAD errors, but not to DOA estimation errors.

While the open-loop cognitive-driven speech enhancement system in Chapter 3 is able to suppress the interfering speaker and background noise, it is not designed to suppress reverberation. In **Chapter 4**, we hence proposed an open-loop cognitive-driven convolutional beamforming system aiming at enhancing the attended speaker and jointly suppressing the interfering speaker, reverberation and background noise. The proposed system consists of a neural-network-based mask estimator, convolutional beamformers and a least-squares-based AAD method which uses a 30-second correlation window. First, the masks of both speakers are estimated from the noisy and reverberant microphone signals using a speech separation neural network. Based on these masks, two convolutional beamformers generate reference signals for AAD by enhancing the speech signal of each speaker. The least-squares-based AAD method then selects one of the reference signals as the enhanced attended speech signal. For the beamformers we proposed to use a wMPDR convolutional beamformer as it combines dereverberation, noise suppression and interfering speaker suppression. We also proposed an extension of the wMPDR convolutional beamformer, referred to as a wLCMP convolutional beamformer, which allows to control the level of suppression of the interfering speaker. We experimentally compared the proposed cognitive-driven convolutional beamforming system with cognitive-driven speech enhancement systems based on (conventional) MVDR, LCMV, MPDR and LCMP beamformers. The experimental results showed that for a two-speaker scenario the proposed system using either the convolutional wMPDR beamformer or

2 Please note that as opposed to Chapter 2, the decoding performance reported in Chapter 3 was obtained using spatio-temporal AAD filters with fixed parameters.

the convolutional wLCMP beamformer yields the highest fwSSNR improvement for the anechoic condition (4.0–4.4 dB[3]) as well as for the reverberant condition (0.5–2.1 dB), whereas the conventional LCMV beamformer yields the highest AAD performance in the reverberant condition ($> 81\%$). In addition, the results showed that the wMPDR convolutional beamformer yields a larger fwSSNR improvement than the wLCMP convolutional beamformer. Although the wMPDR convolutional beamformer strongly suppresses the interfering speaker, it may deprive the listener from the ability to switch attention between the speakers. In contrast, the wLCMP convolutional beamformer is able to both control the interfering speaker suppression as well as yield a considerably large fwSSNR improvement. Furthermore, the experimental results showed that for the considered acoustic setup the AAD performance and the fwSSNR improvement of the proposed system using convolutional beamformers are sensitive to mask estimation errors, particularly for the reverberant condition. Nevertheless, for the reverberant condition only the proposed system using convolutional beamformers provides a fwSSNR improvement, showing the influence of dereverberation, while the systems using conventional beamformers tend to degrade the fwSSNR.

While in Chapters 3 and 4 we considered open-loop cognitive-driven speech enhancement systems, in **Chapter 5** we closed the cognitive-driven speech enhancement system loop by enabling the listener to interact with an adaptive gain controller using a real-time AAD. The real-time AAD infers the probabilistic attention measures of the attended and the unattended speaker from EEG recordings of the listener and the (assumed to be known) speech signals of both speakers. Based on these probabilistic attention measures, the adaptive gain controller amplifies the identified attended speaker and attenuates the identified unattended speaker. The loop of cognitive-driven gain control is then closed by presenting the amplified attended speaker and the attenuated unattended speaker via loudspeakers. The experimental results demonstrated the feasibility of the proposed closed-loop cognitive-driven gain controller system (both using GLM and SSM), enabling the listener to interact with the system in real-time. Although there is a significant delay to detect attention switches (11.5–19.8 seconds), which causes the attended speaker to be wrongly attenuated and the unattended speaker to be wrongly amplified, the proposed closed-loop system is able to improve the SIR between the attended and the unattended speaker (0.6–3.8 dB). In addition, no significant difference in terms of decoding performance and switch detection delay was observed between the proposed closed-loop cognitive-driven system and the open-loop AAD system. Moreover, the subjective evaluation results showed that the proposed closed-loop cognitive-driven system demands a similar perceived level of cognitive effort to follow the attended speaker, to ignore the unattended speaker and to switch attention between both speakers compared to the open-loop AAD system. With this work, a first attempt

3 Please note that due to different reference signals used for the fwSSNR and the SINR, the fwSSNR improvement in Chapter 4 is not directly comparable with the SINR improvement reported in Chapter 3.

was made to bring closed-loop cognitive-driven speech enhancement closer to real-world applications.

While in Chapters 3, 4 and 5 we considered complete open-loop and closed-loop cognitive-driven speech enhancement systems, in Chapters 6 and 7 we focused on AAD and proposed two novel methods to improve the decoding performance.

In **Chapter 6**, we proposed reference signal generation approaches based on binary masking, which discard low-energy intervals susceptible to interfering speech and background noise. The proposed approaches determine the binary masks from the directional speech presence probability and the DOAs of both speakers, which are estimated from the microphone signals. The reference signals for decoding are then generated by either masking the microphone signals or the MVDR output signals. The reference signals comprising moderate-to-high speech energy intervals of the oracle reference signals generated by masking were also considered. Moreover, the binary masks were considered themselves as reference signals for decoding. The performance of the proposed reference signal generation algorithms was evaluated using correlation coefficients obtained with a 30-second correlation window for a two-speaker scenario in diffuse babble noise. The experimental results show that the reference signals comprising moderate-to-high speech energy intervals of the oracle reference signals can be reliably used instead of the (complete) oracle reference signals for decoding auditory attention. This suggests that for generating appropriate reference signals, e.g., from the microphone signals, it is more important to preserve the moderate-to-high speech energy intervals than low speech energy intervals. In addition, the experimental results showed that the proposed reference signal generation algorithms masking either the microphone signals or the MVDR output signals significantly improve the decoding performance ($> 81\%$) compared to when using the (non-masked) microphone signals ($> 77\%$) and the MVDR output signals ($> 78\%$) especially in the reverberant condition. Quite remarkably, the results show that AAD using the binary masks as reference signals yields a decoding performance that is comparable to using the masked MVDR output signals as reference signals.

In **Chapter 7**, we proposed a novel AAD method based on a state-space model, which improves the decoding performance of a linear (least-squares-based) AAD method and a non-linear (DNN-based) AAD method using a short correlation window. First, correlation coefficients are generated by both AAD methods using a 5-second correlation window. Similarly as in Chapter 5, a state-space model then translates the generated correlation coefficients into probabilistic attention measures, based on which the attended speaker is identified. For different acoustic conditions we experimentally compared the performance of the AAD method using the state-space model with the least-squares-based AAD method and the DNN-based AAD method. The results showed that when using either the least-squares-based or the DNN-based AAD method, a relatively low decoding performance (69%–73%) is obtained, mainly due to the relatively small (attended and unattended) correlation coefficients with a large variability based on which decoding is performed. When using the state-space model, the decoding performance significantly increases, where the increase is considerably larger for the least-squares-based AAD method ($> 94\%$) than for the DNN-based AAD method ($> 73\%$).

8.2 Further research directions

In this section, we summarize possible research directions for further improvements and possible applications of the proposed AAD methods and cognitive-driven speech enhancement systems.

While the discussion in Chapter 2 was limited to AAD methods in which auditory attention is decoded based on envelope reconstruction from EEG recordings and a comparison with the envelopes of the reference signals, AAD approaches based on integrated DNN-based architecture as in [130], which directly compare the EEG recordings with the envelopes of the reference signals, or forward models as in [22, 128], which predict the EEG recordings from the envelopes, remain to be investigated. Further work could therefore include a study on how reverberation, noise and interfering speakers influence these AAD approaches. Furthermore, in [231–233] it was shown that phonetic features and spectrograms of speech signals are better encoded in EEG recordings compared to the speech envelope. Therefore, it would certainly be interesting to evaluate the potential of incorporating different speech representations or a combination of speech representations into AAD.

The results in Chapter 3 showed that although the cognitive-driven binaural LCMV beamformer with interference suppression factor $\delta = 0.1$ is able to significantly improve the binaural SINR, the binaural SINR improvement is sensitive to AAD errors. When trials are wrongly decoded, the unattended speaker is wrongly enhanced by the binaural LCMV beamformer and in addition the attended speaker is wrongly suppressed, such that the binaural SINR improvement averaged over wrongly decoded trials can even be negative. It was also shown that for $\delta = 0.2$, the binaural SINR improvement is less prone to wrongly decoded trials compared to smaller values of δ. This suggests that an adaptive interference suppression factor control, e.g., based on the probabilistic attention measures, can effectively improve the sensitivity of the binaural SINR improvement to AAD errors.

In Chapter 4, it was shown that the proposed cognitive-driven convolutional wLCMP beamformer yields the highest fwSSNR improvement for the reverberant condition, whereas the (conventional) LCMV beamformer yields the highest AAD performance for the reverberant condition. Future work could therefore investigate the potential of combining the convolutional and the conventional beamformers to improve both the decoding as well as the speech enhancement performance. Furthermore, it was shown that when using estimated mask-based RTFs instead of oracle-mask-based RTFs for the convolutional beamformers the AAD performance decreases, especially for the reverberant condition. In addition, for the reverberant condition the conventional beamformers using estimated anechoic RTFs selected from a database of prototype RTFs yield a larger AAD performance than the convolutional beamformers using estimated mask-based reverberant RTFs. This suggests that in order to improve the AAD performance, a better estimation of RTFs is required, e.g., based on prototype RTFs, neural networks that are more robust to background noise and reverberation for mask estimation, or a combination of prototype RTFs and neural networks.

The open-loop and closed-loop cognitive-driven speech enhancement systems proposed in this thesis are based on a modular processing flow first decoding auditory attention and then performing speech enhancement. Alternatively, decoding and speech enhancement may be jointly performed, which could lead to an optimal end-to-end cognitive-driven speech enhancement system. In [234] an end-to-end neural-network-based speech enhancement method for extracting a target speaker from the mixture has been proposed which is based on an adaptation utterance spoken by the target speaker to guide the neural networks. Future work could therefore investigate an end-to-end cognitive-driven speech enhancement system using neural networks that are guided by EEG recordings to extract the attended speaker.

In Chapter 5, the closed-loop AAD experiments using the proposed cognitive-driven gain controller system were performed without incorporating a practicing phase for the participants. However, the subjective evaluation results showed that a significant improvement in the system usage was obtained throughout the closed-loop AAD experiment. Future work could therefore investigate the impact of incorporating a practicing phase on the decoding and the speech enhancement performance of the cognitive-driven gain controller system.

In Chapter 6, we proposed reference signal generation algorithms for AAD based on binary masking, which discard low-energy intervals susceptible to interfering speech and background noise. The proposed algorithms determine the binary masks based on the DSPP of both speakers and generate the reference signals by either masking the microphone signals or the MVDR output signals. Although the proposed algorithms are able to discard intervals in the reference signals with a low target speech energy, they may not be able to discard intervals with a low SINR, which are dominated by interfering speech and background noise. The intervals with a low SINR can be determined based on, e.g., the DSPP ratios of both speakers. Further work could therefore include a study on improving the performance of the proposed reference signal generation algorithms by incorporating the DSPP ratios in binary masking.

Although the application of the proposed open-loop and closed-loop cognitive-driven speech enhancement systems in this thesis was limited to acoustic scenarios with two competing speakers, it was shown in [209] that open-loop AAD is feasible for an acoustic scenario with four competing speakers when using perfectly separated clean speech signals for decoding. In addition, the application of the proposed cognitive-driven speech enhancement systems was limited to acoustic scenarios with non-moving speakers, while in real-world conditions speakers may move. Therefore, it would certainly be interesting to investigate the performance of (an extension of) the proposed cognitive-driven speech enhancement systems for acoustic scenarios with more than two competing and moving speakers.

Furthermore, it would be interesting to investigate the potential of combining other signal modalities with EEG for cognitive-driven speech enhancement, e.g., eye gaze direction, eye blinks and pupillometry of the listener or video recordings. The eye gaze direction can be estimated, e.g., from electrooculography (EOG) recordings [235, 236]. The eye blinks, the pupil dilation and the eye gaze direction can be measured, e.g., using an eye tracker camera, mounted on glasses. These mea-

surements can be informative of cognitive processes such as cognitive load [237]. Video recordings can be combined with an acoustic analysis, e.g., to estimate the DOAs and the number of speakers. In addition, the (silent) video recordings of lip movement of speakers can be used to reconstruct the speech of speakers [238–240]. Further work could therefore include a study on improving the performance of the proposed cognitive-driven speech enhancement systems by combining several signal modalities.

Finally, the application of AAD is obviously not limited to hearing devices. AAD could also be used, e.g., for virtual reality (VR) that simulates a remote environment, e.g., for education, training, entertainment and medicine. AAD could be used to adapt the simulated world based on the auditory attention of the VR user. Another application of AAD could be in work environments demanding high alertness to audio inputs, e.g., air/train traffic control stations. AAD could be used in these environments to monitor the state of the auditory attention of an operator in order not to miss salient audio inputs.

BIBLIOGRAPHY

[1] Y. Hochberg and A. C. Tamhane, *Multiple Comparison Procedures*. John Wiley and Sons, 1987.

[2] World Health Organization (WHO), "Addressing the rising prevalence of hearing loss," *World Health Organization*, 2018. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/260336/9789241550260-eng.pdf?sequence=1&isAllowed=y

[3] H. Wahl and V. Heyl, "Connections between vision, hearing, and cognitive function in old age," *Generations*, vol. 27, no. 1, pp. 39–45, 2000.

[4] G. Gates and J. Mills, "Presbycusis," *The Lancet*, Feb. 2005.

[5] P. B. Baltes and U. Lindenberger, "Emergence of a powerful connection between sensory and cognitive functions across the adult life span," *Psychology and Aging*, vol. 12, no. 1, pp. 12–21, 1997.

[6] F. R. Lin, K. Yaffe, J. Xia, Q.-L. Xue, T. B. Harris, E. Purchase-Helzner, S. Satterfield, H. N. Ayonayon, L. Ferrucci, and E. M. Simonsick, "Hearing loss and cognitive decline in older adults," *JAMA Internal Medicine*, vol. 173, no. 4, pp. 293–299, 2013.

[7] H. Dillon, *Hearing Aids*. Thieme Medical Publishers Inc, 2012.

[8] K. R. Borisagar, R. M. Thanki, and B. S. Sedani, *Speech Enhancement Techniques for Digital Hearing Aids*. Springer International Publishing, 2019.

[9] B. D. van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[10] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[11] D. Marquardt, "Development and evaluation of psychoacoustical ly motivated binaural noi se reduction and cue preservation techniques," Ph.D. dissertation, Ph.D. dissertation, University of Oldenburg, Oldenburg, Germany, Nov. 2015.

[12] E. Vincent, T. Virtanen, and S. Gannot, *Auditory Scene Analysis*. Wiley, 2017.

[13] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[14] M. Taseska, "Informed spatial filters for speech enhancement noise and interference reduction, blind source separation, and acoustic source tracking," Ph.D. dissertation, Ph.D. dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, Nov. 2017.

[15] R. Rehr, "Robust speech enhancement using statistical signal processing and machine-learning," Ph.D. dissertation, Ph.D. dissertation, University of Hamburg, Hamburg, Germany, Apr. 2018.

[16] 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*, 2014.

[17] C. Horton, R. Srinivasan, and M. D'Zmura, "Envelope responses in single-trial EEG indicate attended speaker in a cocktail party," *Neural Engineering*, vol. 11, no. 4, p. 46015, 2014.

[18] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of Neural Engineering*, vol. 12, no. 4, p. 46007, 2015.

[19] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Auditory attention decoding with EEG recordings using noisy acoustic reference signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 694–698.

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

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

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

[23] 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*, Dec. 2018.

[24] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, "Real-time tracking of selective auditory attention from M/EEG: A Bayesian filtering approach," *Frontiers in Neuroscience*, vol. 12, p. 262, 2018.

[25] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of Neural Engineering*, vol. 15, no. 6, p. 66017, 2018.

[26] A. Aroudi, B. Mirkovic, M. De Vos, and S. Doclo, "Impact of different acoustic components on EEG-based auditory attention decoding in noisy and reverberant conditions," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 652–663, April 2019.

[27] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1045–1056, 2017.

[28] J. O'Sullivan, Z. Chen, J. Herrero, G. M. McKhann, S. A. Sheth, A. D. Mehta, and N. Mesgarani, "Neural decoding of attentional selection in multi-speaker environments without access to clean sources," *Journal of Neural Engineering*, vol. 14, no. 5, p. 56001, 2017.

[29] A. Aroudi, D. Marquardt, and S. Doclo, "EEG-based auditory attention decoding using steerable binaural superdirective beamformer," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 851–855.

[30] W. Pu, J. Xiao, T. Zhang, and Z. Luo, "A joint auditory attention decoding and adaptive binaural beamforming algorithm for hearing devices," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2019, pp. 311–315.

[31] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 862–875, 2020.

[32] A. Aroudi and S. Doclo, "EEG-based auditory attention decoding: Impact of reverberation, noise and interference reduction," in *Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Banff, Canada, Oct. 2017, pp. 3042–3047.

[33] S. Gordon-Salant and P. J. Fitzgibbons, "Recognition of multiply degraded speech by young and elderly listeners," *Journal of Speech and Hearing Research*, vol. 38, no. 5, pp. 1150–1156, 1995.

[34] S. Gordon-Salant, G. H. Yeni-Komshian, P. J. Fitzgibbons, and J. Barrett, "Age-related differences in identification and discrimination of temporal cues in speech segments," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2455–2466, 2006.

[35] S. Gordon-Salant, G. H. Yeni-Komshian, P. J. Fitzgibbons, and J. I. Cohen, "Effects of age and hearing loss on recognition of unaccented and accented multisyllabic words," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 884–897, 2015.

[36] M. A. Wardenga, N.and Zokoll, B. Kollmeier, and H. Maier, "Influence of background noise level on speech reception in normal and hearing impaired persons," *Laryngo-Rhino-Otol*, vol. 97, no. S 02, pp. 10 496–10 625, 2018.

[37] H. Kuttruff, *Room acoustics*. Taylor and Francis, 2000.

[38] J. S. Bradley, H. Sasto, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, Jun. 2003.

[39] A. Warzybok, J. Rennies, T. Brand, S. Doclo, and B. Kollmeier, "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 269–282, Jan 2013.

[40] I. Arweiler and J. M. Buchholz, "The influence of spectral characteristics of early reflections on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 130, no. 2, pp. 996–1005, Aug. 2011.

[41] A. K. Nàbělek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1259–1265, 1989.

[42] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, 2006.

[43] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer, 2010.

[44] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, 1992.

[45] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. Cambridge, Mass. MIT Press, 1997.

[46] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 833–843, 2004.

[47] V. Best, S. Carlile, N. Kopčo, and A. van Schaik, "Localization in speech mixtures by listeners with hearing loss," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. EL210–EL215, 2011.

[48] M. Lavandier and J. F. Culling, "Speech segregation in rooms: Effects of reverberation on both target and interferer," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1713–1723, 2007.

[49] A. K. C. Lee and B. G. Shinn-Cunningham, "Effects of reverberant spatial cues on attention-dependent object formation," *Journal of the Association for Research in Otolaryngology*, vol. 9, no. 1, pp. 150–160, 2008.

[50] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[51] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*. Morgan and Claypool, 2013.

[52] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. USA: CRC Press, Inc., 2013.

[53] J. Benesty, *Fundamentals of Speech Enhancement*. Springer, 2019.

[54] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[55] R. C. Hendriks, R. Heusdens, and J. Jensen, "Log-spectral magnitude MMSE estimators under super-gaussian densities," in *Proceedings of INTERSPEECH*, Brighton, United Kingdom, Jan. 2009, pp. 1319–1322.

[56] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Audio, Speech, and*

Language Processing, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[57] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 7, pp. 1–17, 2005.

[58] A. Aroudi, H. Veisi, and H. Sameti, "Hidden markov model-based speech enhancement using multivariate laplace and gaussian distributions," *IET Signal Processing*, vol. 9, no. 2, pp. 177–185, 2015.

[59] Joon-Hyuk Chang, Nam Soo Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[60] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[61] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise psd tracking with low complexity," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2010, pp. 4266–4269.

[62] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using mmse domain subspace decompositions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 541–553, Mar. 2008.

[63] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[64] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori snr estimation approach based on selective cepstro-temporal smoothing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2008, pp. 4897–4900.

[65] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[66] H. Sameti, H. Sheikhzadeh, Li Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 445–455, Sep. 1998.

[67] F. Deng, C. Bao, and W. B. Kleijn, "Sparse hidden markov models for speech enhancement in non-stationary noise environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1973–1987, Nov. 2015.

[68] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 253–256, Mar. 2013.

[69] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[70] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1233–1242, Jul. 2015.

[71] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 998–1011, May 2013.

[72] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[73] ——, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[74] Q. He, F. Bao, and C. Bao, "Multiplicative update of auto-regressive gains for codebook-based speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 457–468, Mar. 2017.

[75] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.

[76] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[77] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2018, pp. 696–700.

[78] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," in *Proceedings of INTERSPEECH*, HYDERABAD, INDIA, Sep. 2018, pp. 342–346.

[79] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal timeâ€"frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[80] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[81] S. E. Chazan, J. Goldberger, and S. Gannot, "A hybrid approach for speech enhancement using MoG model and neural network phoneme classifier," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2516–2530, Dec. 2016.

[82] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016.

[83] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[84] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, oct. 2018.

[85] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "A unified framework for neural speech separation and extraction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2019, pp. 6975–6979.

[86] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays*. Springer, 2001, pp. 19–38.

[87] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, no. 3, pp. 229–240, Jun. 1996.

[88] J. Benesty, J. Chen, and C. Pan, *Fundamentals of Differential Beamforming*. Springer, 2016.

[89] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

[90] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2014, Juan-les-Pins, France, pp. 99–103.

[91] N. Gößling, D. Marquardt, I. Merks, T. Zhang, and S. Doclo, "Optimal binaural LCMV beamforming in complex acoustic scenarios: Theoretical and practical insights," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, Tokyo, Japan, pp. 381–385.

[92] S. Doclo and M. Moonen, "Gsvd-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[93] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, Feb. 2010.

[94] L. W. Brooks and I. S. Reed, "Equivalence of the likelihood ratio processor, the maximum signal-to-noise ratio filter, and the Wiener filter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 8, no. 5, pp. 690–692, Sep. 1972.

[95] B. Cornelis, S. Doclo, T. van den Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 342–355, 2010.

[96] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel wiener filtering techniques for noise reduction," *Speech Enhancement*, 2005.

[97] ——, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7, pp. 636–656, 2007.

[98] T. Nakatani and K. Kinoshita, "Maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Sep. 2019, pp. 1–5.

[99] T. Nakatani, K. Kinoshita, R. Ikeshita, H. Sawada, and S. Araki, "Simultaneous denoising, dereverberation, and source separation using a unified convolutional beamformer," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2019, New Paltz, NY, USA, pp. 224–228.

[100] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly optimal dereverberation and beamforming," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May. 2020.

[101] N. Ito, C. Schymura, S. Araki, and T. Nakatani, "Noisy cGMM: Complex Gaussian mixture model with non-sparse noise model for joint source separation and denoising," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 1662–1666.

[102] S. Chakrabarty and E. A. P. Habets, "Timeâ€"frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, Aug. 2019.

[103] F. Bahmaninezhad, J. Y. Wu, R. Gu, S.-X. Zhang, Y. Xu, and M. Yu, "A comprehensive study of speech separation: Spectrogram vs waveform separation," in *Proceedings of INTERSPEECH*, Graz, Austria, Sep. 2019.

[104] R. Serizel, M.Moonen, B.VanDijk, and J.Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, p. 785â€"799, 2014.

[105] I. Cohen,, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[106] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD,

Australia, Apr. 2015, pp. 544–548.

[107] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[108] N. Ito and S. Araki, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2017, pp. 681–685.

[109] M. Jeub, M. Dörbecker, and P. Vary, "A semi-analytical model for the binaural coherence of noise fields," *IEEE Signal Processing Letters*, vol. 18, no. 3, pp. 197–200, Mar. 2011.

[110] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 6, 2009.

[111] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, *Handbook on Array Processing and Sensor Networks (S. Haykin, K. J. Ray Liu, eds.)*. Wiley, 2010, ch. Acoustic beamforming for hearing aid applications, pp. 269–302.

[112] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2449–2464, Dec. 2015.

[113] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 484–497, 2008.

[114] A. W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1508–1516, 1988.

[115] S. Doclo, S. Gannot, D. Marquardt, and E. Hadad, "Binaural speech processing with application to hearing devices," in *Audio Source Separation and Speech Enhancement*. Wiley, 2019.

[116] A. H. Kamkar-Parsi and M. Bouchard, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 521–533, May 2009.

[117] ——, "Instantaneous binaural target psd estimation for hearing aid noise reduction in complex acoustic environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 4, pp. 1141–1154, Apr. 2011.

[118] M. Jeub, C. Nelke, H. Krüger, C. Beaugeant, and P. Vary, "Robust dual-channel noise power spectral density estimation," *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 2304–2308, Aug. 2011.

[119] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communication*, vol. 27, no. 3, pp. 209–222, 1999.

[120] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[121] L. W. Brooks and I. S. Reed, "Equivalence of the likelihood ratio processor, the maximum signal-to-noise ratio filter, and the Wiener filter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-8, no. 5, pp. 690–692, Sep. 1972.

[122] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 543–558, Mar. 2016.

[123] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Theoretical analysis of linearly constrained multi-channel Wiener filtering algorithms for combined noise reduction and binaural cue preservation in binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2384–2397, Dec. 2015.

[124] D. Marquardt and S. Doclo, "Interaural coherence preservation for binaural noise reduction using partial noise estimation and spectral postfiltering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1261–1274, Jul. 2018.

[125] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1579–1585, Apr. 2007.

[126] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in MWF-based binaural noise reduction algorithms using partial noise estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2015, pp. 654–658.

[127] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van de Par, "Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, pp. 1–11, 2016.

[128] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Frontiers in Neuroscience*, vol. 13, p. 153, 2019.

[129] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, May 2017.

[130] 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 meth-

ods," *Scientific Reports, Nature*, vol. 9, no. 11538, Aug. 2019.

[131] 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*, pp. 1–1, 2020.

[132] 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," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/08/11/475673

[133] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mTRF) toolbox: A matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in Human Neuroscience*, vol. 10, p. 604, 2016.

[134] B. Graimann, B. Z. Allison, and G. Pfurtscheller, *Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction*. Springer, 2013.

[135] B. Allison, S. Dunne, R. Leeb, J. Del R. Millàn, and A. Nijholt, *Towards Practical Brain-Computer Interfaces: Bridging the Gap from Research to Real-World Applications*. Springer, 2013.

[136] D. P. Subha, P. K. Joseph, R. Acharya U, and C. M. Lim, "EEG signal analysis: A survey," *Journal of Medical Systems*, vol. 34, no. 2, pp. 195–212, Apr. 2010.

[137] D. L. Schomer and F. H. L. da Silva, *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Oxford, UK: Oxford University Press, Nov. 2017.

[138] L. Bi, X. Fan, and Y. Liu, "EEG-based brain-controlled mobile robots: A survey," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 2, pp. 161–176, Mar. 2013.

[139] M. Finke, A. Büchner, E. Ruigendijk, M. Meyer, and P. Sandmann, "On the relationship between auditory cognition and speech intelligibility in cochlear implant users: An erp study," *Neuropsychologia*, vol. 87, pp. 169 – 181, 2016.

[140] D. Bansal and R. Mahajan, *EEG-Based Brain-Computer Interfaces: Cognitive Analysis and Control Applications*. Academic Press, 2019.

[141] L. Jacquemin, G. Mertens, W. Schlee, P. V. de Heyning, and A. Gilles, "Literature overview on P3 measurement as an objective measure of auditory performance in post-lingually deaf adults with a cochlear implant," *International Journal of Audiology*, pp. 1–8, 2019.

[142] M. Hassani and M. R. Karami, "Noise estimation in electroencephalogram signal by using volterra series coefficients," *Journal of medical signals and sensors*, vol. 5, pp. 192–200, 2015.

[143] J. A. Urigüen and B. Garcia-Zapirain, "EEG artifact Removal-State-Of-The-Art and guidelines," *Journal of Neural Engineering*, vol. 12, no. 3, p. 031001, 2015.

[144] T. Handy, *Event-Related Potentials: A Methods Handbook*. Cambridge, USA: The MIT Press, 2005.

[145] T. Picton, "Hearing in time: Evoked potential studies of temporal processing," *Ear and Hearing*, vol. 34, no. 4, pp. 385–401, 2013.

[146] G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger, "Linear systems analysis of functional magnetic resonance imaging in human v1," *Journal of Neuroscience*, vol. 16, no. 13, pp. 4207–4221, 1996.

[147] D. Ringach and R. Shapley, "Reverse correlation in neurophysiology," *Cognitive Science*, vol. 28, no. 2, pp. 147–166, 2004.

[148] C. R. Holdgraf, J. W. Rieger, C. Micheli, S. Martin, R. T. Knight, and F. E. Theunissen, "Encoding and decoding models in cognitive electrophysiology," *Frontiers in Systems Neuroscience*, vol. 11, p. 61, 2017.

[149] S. V. David, N. Mesgarani, and S. A. Shamma, "Estimating sparse spectro-temporal receptive fields with natural stimuli," *Network: Computation in Neural Systems*, vol. 18, no. 3, pp. 191–212, 2007.

[150] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European Journal of Neuroscience*, vol. 31, no. 1, pp. 189–193, 2010.

[151] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Phoneme representation and classification in primary auditory cortex," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 899–909, 2008.

[152] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 854–11 859, 2012.

[153] W. Bialek, F. Rieke, R. de Ruyter van Steveninck, and D. Warland, "Reading a neural code," *Science*, vol. 252, no. 5014, pp. 1854–1857, 1991.

[154] F. Rieke, D. A. Bodnar, and W. Bialek, "Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 262, no. 1365, pp. 259–265, 1995.

[155] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96 – 110, 2014.

[156] B. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS Biology*, vol. 10, no. 1, p. 1001251, 2012.

[157] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of Neurophysiology*, vol. 107, no. 1, pp. 78–89, 2011.

[158] ——, "Cortical entrainment to continuous speech: functional roles and interpretations," *Frontiers in human neuroscience*, vol. 8, p. 311, 2014.

[159] 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*, vol. 102, no. 6, pp. 3329–3339, 2009.

[160] N. Ding and J. Z. Simon, "Adaptive temporal encoding leads to a background-insensitive cortical representation of speech," *Journal of Neuroscience*, vol. 33, no. 13, pp. 5728–5735, Mar. 2013.

[161] Y.-Y. Kong, A. Mullangi, and N. Ding, "Differential modulation of auditory responses to attended and unattended speech in different listening conditions," *Hearing Research*, vol. 316, pp. 73 – 81, 2014.

[162] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N. E. Crone, J. Rieger, G. Schalk, R. T. Knight, and B. N. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in Neuroengineering*, vol. 7, p. 14, 2014.

[163] R. Zink, A. Baptist, A. Bertrand, S. Van Huffel, and M. De Vos, "Online detection of auditory attention in a neurofeedback application," in *Proc. 8th International Workshop on Biosignal Interpretation*, Osaka, Japan, Nov. 2016, pp. 1–4.

[164] N. Das, W. Biesmans, A. Bertrand, and T. Francart, "The effect of head-related filtering and ear-specific decoding bias on auditory attention detection," *Journal of Neural Engineering*, vol. 13, no. 5, p. 056014, 2016.

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

[166] N. Das, S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-based attention-driven speech enhancement for noisy speech mixtures using N-fold multi-channel Wiener filters," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Kos, Greece, Aug. 2017, pp. 1660–1664.

[167] C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, and N. Mesgarani, "Speaker-independent auditory attention decoding without access to clean speech sources," *Science Advances*, vol. 5, no. 5, 2019.

[168] N. Das, J. Zegers, H. V. hamme, T. Francart, and A. Bertrand, "Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding," *Journal of Neural Engineering*, vol. 17, no. 4, p. 046039, Aug. 2020.

[169] A. Aroudi and S. Doclo, "EEG-based auditory attention decoding using unprocessed binaural signals in reverberant and noisy conditions," in *Proc. Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jeju, South Korea, 2017, pp. 484–488.

[170] ——, "Cognitive-driven binaural LCMV beamformer using EEG-based auditory attention decoding," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019, pp. 406–410.

[171] A. Aroudi, M. Delcroix, T. Nakatani, K. Kinoshita, S. Araki, and S. Doclo, "Cognitive-driven convolutional beamforming using EEG-based auditory attention decoding," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Espoo, Finland, Sep. 2020, pp. 1–6.

[172] A. Aroudi, T. de Taillez, and S. Doclo, "Improving auditory attention decoding performance of linear and non-linear methods using state-space model," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 8703–8707.

[173] J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay, *The Auditory System at the Cocktail Party*. Springer, 2017.

[174] B. G. Shinn-Cunningham and V. Best, "Selective attention in normal and impaired hearing," *Trends in Amplification*, vol. 12, no. 4, pp. 283–299, 2008.

[175] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, May 2012.

[176] E. B. Petersen, M. Wöstmann, J. Obleser, and T. Lunner, "Neural tracking of attended versus ignored speech is differentially affected by hearing loss," *Journal of neurophysiology*, vol. 117, no. 1, pp. 18–27, 2017.

[177] N. Ding, M. Chatterjee, and J. Z. Simon, "Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure," *Neuroimage*, vol. 88, pp. 41–46, 2014.

[178] P. Patel, L. K. Long, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Joint representation of spatial and phonetic features in the human core auditory cortex," *Cell reports*, vol. 24, no. 8, pp. 2051–2062, 2018.

[179] L. Fiedler, M. Woestmann, 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*, vol. 14, no. 3, p. 36020, 2017.

[180] R. Zink, S. Proesmans, A. Bertrand, S. Van Huffel, and M. De Vos, "Online detection of auditory attention with mobile EEG: closing the loop with neurofeedback," *bioRxiv*, 2017.

[181] N. Das, S. Van Eyndhoven, T. Francart, and A. Bertrand, "Adaptive attention-driven speech enhancement for EEG-informed hearing prostheses," in *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, Florida (USA), Aug. 2016, pp. 77–80.

[182] B. Ekin, L. Atlas, M. Mirbagheri, and A. K. C. Lee, "An alternative approach for auditory attention tracking using single-trial EEG," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 20-25 March 2016, pp. 729–733.

[183] C. Darwin and R. Hukin, "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention," *The Journal of the Acoustical Society of America*, vol. 108, no. 1, pp. 335–342, 2000.

[184] J. F. Culling, K. I. Hodder, and C. Y. Toh, "Effects of reverberation on perceptual segregation of competing voices," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2871–2876, 2003.

[185] P. M. Zurek, R. L. Freyman, and U. Balakrishnan, "Auditory target detection in reverberation," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1609–1620, 2004.

[186] D. Ruggles and B. Shinn-Cunningham, "Spatial selective auditory attention in the presence of reverberant energy: individual differences in normal-hearing listeners," *Journal of the Association for Research in Otolaryngology*, vol. 12, no. 3, pp. 395–405, 2011.

[187] S. Anderson, E. Skoe, B. Chandrasekaran, and N. Kraus, "Neural timing is linked to speech perception in noise," *Journal of Neuroscience*, vol. 30, no. 14, pp. 4922–4926, 2010.

[188] H. Ohrka, "Ohrka.de-kostenlose Hörabenteuer für Kinderohren," 2012. [Online]. Available: http://www.ohrka.de

[189] E. Hering, "Kostbarkeiten aus dem Deutschen Märchenschatz," *Audiopool Hörbuchverlag MP3 CD*, ISBN: 9783868471175, 2011.

[190] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. of International Conference on Digital Signal Processing (DSP)*, Santorini, Greece, Jul. 2009, pp. 1–5.

[191] J. Thiemann and S. Van De Par, "Multiple model high-spatial resolution HRTF measurements," in *Proc. DAGA*, 2015.

[192] E. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

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

[194] D. Looney, C. Park, Y. Xia, P. Kidmose, M. Ungstrup, and D. P. Mandic, "Towards estimating selective auditory attention from EEG using a novel time-frequency-synchronisation framework," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, July 2010, pp. 1–5.

[195] O. Etard, M. Kegler, C. Braiman, A. E. Forte, and T. Reichenbach, "Real-time decoding of selective attention from the human auditory brainstem response to continuous speech," *bioRxiv*, 2018.

[196] Z. Cao and C. Lin, "Inherent fuzzy entropy for the improvement of EEG complexity evaluation," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 2, pp. 1032–1035, April 2018.

[197] S. Akram, J. Z. Simon, and B. Babadi, "Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1896–1905, 2017.

[198] T. Dau, J. Maercher Roersted, S. Fuglsang, and J. Hjortkjær, "Towards cognitive control of hearing instruments using EEG measures of selective attention," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1744–1744, 2018.

[199] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE/ACM Transactions on Audio, Speech,*

and Language Processing, vol. 20, no. 7, pp. 2016–2030, Sep. 2012.

[200] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[201] J. Chen, J. Benesty, and Y. Huang, "Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 1, pp. 25–36, Jan. 2005.

[202] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1507–1519, Oct 2019.

[203] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2015, New Paltz, NY, USA, pp. 1–5.

[204] D. Marquardt and S. Doclo, "Noise power spectral density estimation for binaural noise reduction exploiting direction of arrival estimates," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2017, New Paltz, NY, USA, pp. 234–238.

[205] E. Habets, J. Benesty, and P. A. Naylor, "A speech distortion and interference rejection constraint beamformer," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 854–867, Mar. 2012.

[206] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, Aug 1972.

[207] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, Mar. 2017, pp. 11–15.

[208] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, no. 5, pp. 592–605, 2011, perceptual and Statistical Audition.

[209] P. J. Schäfer, F. I. Corona-Strauss, R. Hannemann, S. A. Hillyard, and D. J. Strauss, "Testing the limits of the stimulus reconstruction approach: Auditory attention decoding in a four-speaker free field environment," *Trends in Hearing*, vol. 22, pp. 1–12, Jan 2018.

[210] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 626–630.

[211] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 5206–5210.

[212] C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, and N. Mesgarani, "Speaker-independent auditory attention decoding without access to clean speech sources," *Science Advances*, vol. 5, no. 5, 2019.

[213] 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*, vol. 28, no. 1, pp. 307–317, 2020.

[214] D. Collett, *Modeling Binary Data*. New York: Chapman and Hall, 2002.

[215] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society*, vol. 135, no. 3, pp. 370–384, 1972.

[216] K. E. Train, *Discrete Choice Methods with Simulation*, 2nd ed. Cambridge University Press, 2009.

[217] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, Massachusetts: The MIT Press, 1990.

[218] N. Marrone, C. R. Mason, and G. Kidd, "The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3064–3075, 2008.

[219] M. W. Hoffman, Z. Li, and D. Khataniar, "GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 175–178, 2001.

[220] S. Srinivasan and K. Janse, "Spatial audio activity detection for hearing aids," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, United States, April 2008, pp. 4021–4024.

[221] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, 2010.

[222] S. Meier and W. Kellermann, "Artificial neural network-based feature combination for spatial voice activity detection," in *Proceedings of INTERSPEECH*, San Francisco, US, Sep. 2016, pp. 2987–2991.

[223] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgre, and V. Zue, "TIMIT acousticphonetic continuous speech corpus," 1993.

[224] K. B. Doelling, L. H. Arnal, O. Ghitza, and D. Poeppel, "Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing," *NeuroImage*, vol. 85, pp. 761–768, 2014, new Horizons for Neural Oscillations.

[225] S. J. Kayser, R. A. Ince, J. Gross, and C. Kayser, "Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha," *Journal of Neuroscience*, vol. 35, no. 44, pp. 14 691–14 701, 2015.

[226] Y. Oganian and E. F. Chang, "A speech envelope landmark for syllable encoding in human superior temporal gyrus," *bioRxiv*, 2018. [Online].

Available: https://www.biorxiv.org/content/early/2018/08/09/388280

[227] O. Ghitza, "On the role of theta-driven syllabic parsing in decoding speech: Intelligibility of speech with a manipulated modulation spectrum," *Frontiers in Psychology*, vol. 3, p. 238, 2012.

[228] T. Dozat, "Incorporating nesterov momentum into adam," in *International Conference on Learning Representations (ICLR 2016 workshop)*, 2016.

[229] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[230] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[231] G. M. D. Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.

[232] D. Lesenfants, J. Vanthornhout, E. Verschueren, L. Decruy, and T. Francart, "Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations," *Hearing Research*, vol. 380, pp. 1–9, 2019.

[233] D. Lesenfants, J. Vanthornhout, E. Verschueren, and T. Francart, "Data-driven spatial filtering for improved measurement of cortical tracking of multiple representations of speech," *Journal of Neural Engineering*, vol. 16, no. 6, p. 066017, 2019.

[234] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, aug. 2019.

[235] A. Favré-Felix, R. Hietkamp, C. Graversen, T. Dau, and T. Lunner, "Steering of audio input in hearing aids by eye gaze through electrooculography," *Trends in Hearing*, vol. 6, pp. 135–142, Feb. 2018.

[236] G. Grimm, H. Kayser, M. Hendrikse, and V. Hohmann, "A gaze-based attention model for spatially-aware hearing aids," in *Speech Communication; 13th ITG-Symposium*, Apr. 2018, pp. 1–5.

[237] D. Wendt, R. Hietkamp, and T. Lunner, "Impact of noise and noise reduction on processing effort: A pupillometry study," *Ear and Hearing*, vol. 38, no. 6, pp. 690–700, Dec. 2017.

[238] A. Ephrat and S. Peleg, "Vid2speech: Speech reconstruction from silent video," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2017, pp. 5095–5099.

[239] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 2516–2520.

[240] Y. Li, Z. Liu, Y. Na, Z. Wang, B. Tian, and Q. Fu, "A visual-pilot deep fusion for target speech separation in multitalker noisy environment," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal*

Processing (ICASSP), Barcelona, Spain, May 2020, pp. 4442–4446.

LIST OF PUBLICATIONS

The following publications are related to the work in this thesis.

Peer-reviewed Journal Papers

[J2] A. Aroudi, S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 862-875, Jan. 2020.

[J1] A. Aroudi, B. Mirkovic, M. De Vos, S. Doclo, "Impact of different acoustic components on EEG-based auditory attention decoding in noisy and reverberant conditions," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 652-63, Apr. 2019.

Peer-reviewed Conference Papers

[C7] A. Aroudi, M. Delcroix, T. Nakatani, K. Kinoshita, S. Araki, S. Doclo, "Cognitive-driven convolutional beamforming using EEG-based auditory attention decoding," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Espoo, Finland, pp. 1-6, Sep. 2020.

[C6] A. Aroudi, T. de Taillez, S. Doclo, "Improving auditory attention decoding performance of linear and non-linear methods using state-space model," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 8703-8707, May 2020.

[C5] A. Aroudi, S. Doclo, "Cognitive-driven binaural LCMV beamformer using EEG-based auditory attention decoding," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, pp. 406-410, May 2019.

[C4] A. Aroudi, D. Marquardt, S. Doclo, "EEG-based auditory attention decoding using steerable binaural superdirective beamformer," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, pp. 851-855, Apr. 2018.

[C3] A. Aroudi, S. Doclo, "EEG-based auditory attention decoding: impact of reverberation, noise and interference reduction," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Banff, Canada, pp. 3042-3047, Oct. 2017.

[C2] A. Aroudi, S. Doclo, " EEG-based auditory attention decoding using unprocessed binaural signals in reverberant and noisy conditions," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jeju, South Korea, pp. 484-488, Jul. 2017.

[C1] A. Aroudi, B. Mirkovic, M. De Vos, S. Doclo, "Auditory attention decoding with EEG recordings using noisy acoustic reference signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, pp. 694-698, Mar. 2016.

Papers Submitted or in Submission

[S2] A. Aroudi, E. Fischer, M. Serman, H. Puder, S. Doclo, "Closed-loop cognitive-driven gain control of competing sounds using auditory attention decoding," has been submitted to *Journal of Neural Engineering*.

[S1] A. Aroudi, H. Kayser, S. Doclo, "EEG-based auditory attention decoding using binary masking," to be submitted to *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Peer-reviewed Abstracts

[A8] A. Aroudi, H. Kayser, S. Doclo, "Binary-masking-based auditory attention decoding without access to clean speech signals," in *Auditory EEG Signal Processing (AESoP) Symposium*, Leuven, Belgium, Sep. 2019.

[A7] A. Aroudi, S. Doclo, "Cognitive-driven binaural speech enhancement system for hearing aid applications," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, Jul. 2019.

[A6] A. Aroudi, D. Marquardt, S. Doclo, "Cognitive-driven binaural speech enhancement system for hearing aid applications," in *International Hearing Aid Research Conference (IHCON)*, Lake Tahoe, USA, Aug. 2018.

[A5] A. Aroudi, S. Doclo, "Cognitive-driven binaural speech enhancement system for hearing aid applications," in *6th Summer School of the Joint Research Academy Cluster of Excellence Hearing4all*, Soltau, Germany, Jun. 2018.

[A4] A. Aroudi, D. Marquardt, S. Doclo, "Cognitive-driven binaural speech enhancement system for hearing aid applications," in *Auditory EEG Signal Processing (AESoP) Symposium*, Leuven, Belgium, May 2018.

[A3] A. Aroudi, S. Doclo, "Auditory Attention Decoding in Reverberant and Noisy Conditions," in *Workshop on Signal and Noise along the Auditory Pathway (SNAP)*, Lübeck, Germany, Dec. 2017.

[A2] A. Aroudi, S. Doclo, "Influence of Noisy Reference Signals on Selective Attention Decoding," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Milan, Italy, Aug. 2015.

[A1] A. Aroudi, B. Mirkovic, M. De Vos, S. Doclo, "Auditory attention decoding using noisy reference signals," in *Closing the Auditory (Efferent) Loop International Symposium*, Hanover, Germany, Oct. 2015.