Cognitive-Driven Binaural Beamforming Using EEG-Based Auditory Attention Decoding

Ali Aroudi, *Student Member, IEEE*, and Simon Doclo, *Senior Member, IEEE*

***Abstract*—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 article, 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.**

***Index Terms*—Auditory attention decoding (AAD), steerable binaural beamformer, speech enhancement, direction-of-arrival estimation, EEG signal, brain computer interface.**

I. 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 [1], [2], 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 [3], [4]. 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 [3], neural networks [5], and Bayesian filtering [6]. 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 [6]–[14], closing the loop of an AAD system by presenting feedback to the listener [15], and steering source separation and noise reduction algorithms based on AAD [16]–[19].

The least-squares-based AAD method proposed in [3] 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 [5], [6] 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 [9], [14] 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 [16]–[20] have been proposed. The single-microphone source separation algorithm proposed in [16] 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 [16] show that the algorithm is able to significantly improve

Manuscript received May 22, 2019; revised October 18, 2019 and December 28, 2019; accepted January 11, 2020. Date of publication January 27, 2020; date of current version March 6, 2020. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 390895286 EXC 2177/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Augusto Sarti. *(Corresponding author: Ali Aroudi.)*

The authors are with the Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, 26129 Oldenburg, Germany (e-mail: ali.aroudi@uni-oldenburg.de; simon.doclo@uni-oldenburg.de).

Digital Object Identifier 10.1109/TASLP.2020.2969779

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 [17]–[20] 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 [17], [18] 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 [17], [18] 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 [1], [21], 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 [20] 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. 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 [20], 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.

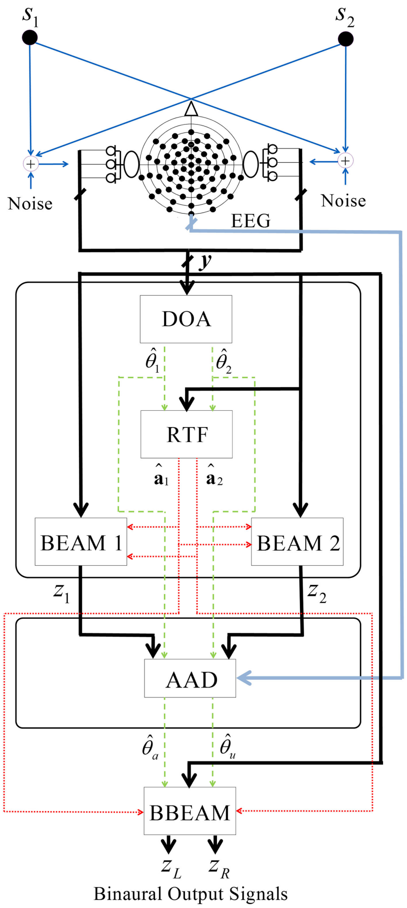


Fig. 1. Block diagram of the proposed cognitive-driven binaural beamformer in an acoustic scenario comprising two competing speakers (s_1 and s_2 with DOAs θ_1 and θ_2) and background noise. Based on the estimated DOAs ($\hat{\theta}_1$ and $\hat{\theta}_2$) of the speakers, the anechoic or reverberant 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.

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

II. 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. 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

$$y_{L,m}[n] = x_{1,L,m}[n] + x_{2,L,m}[n] + v_{L,m}[n], \qquad (1)$$

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 (1).

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

$$\mathbf{y}(k,l) = [Y_{L,1}(k,l) \; \ldots \; Y_{L,M}(k,l) \; Y_{R,1}(k,l) \; \ldots \; Y_{R,M}(k,l)]^T, \qquad (2)$$

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 (1) and (2), the signal vector $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{v}, \qquad (3)$$

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

$$\mathbf{x}_1 = \mathbf{h}_1 S_1, \; \mathbf{x}_2 = \mathbf{h}_2 S_2, \qquad (4)$$

where $\mathbf{h}_1$ and $\mathbf{h}_2$ denote the 2 M-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

$$\mathbf{x}_1 = \bar{\mathbf{a}}_{1,L} X_{1,L,1} = \bar{\mathbf{a}}_{1,R} X_{1,R,1}, \qquad (5)$$

where $\bar{\mathbf{a}}_{1,\{L,R\}}$ denote the 2 M-dimensional relative transfer function (RTF) vectors [1], [2] 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. 1 are obtained by filtering and summing the microphone signals on both hearing aids, i.e.,

$$z = \text{ISTFT}\left\{\mathbf{w}^H \mathbf{y}\right\}, \qquad (6)$$

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

III. 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 IV). Section III-A describes a classification-based DOA estimation algorithm. Section III-B describes two DOA-based RTF vector estimation algorithms.

A. 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 [22], by using a beamforming-based approach [23], or by using a classification-based method [24]. In this paper, we will use the DOA estimation algorithm from [24], 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) [25] is used, which has been shown to be relatively robust to noise and reverberation [26].

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.,

$$\overline{p}_\theta[n] = \tau p_\theta[n] + (1-\tau)\,\overline{p}_\theta[n-1], \qquad (7)$$

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 V), 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 [24] and (7).

B. 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., [27], or selected from a database of (measured) prototype RTF vectors, e.g., [28]. 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 [29]–[32], 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 [33]

$$\tilde{\mathbf{a}}_{\{L,R\}}(k,l) = \frac{\hat{\boldsymbol{\Phi}}_{\mathbf{y}}(k,l)\mathbf{e}_{\{L,R\}}}{\mathbf{e}^T_{\{L,R\}}\hat{\boldsymbol{\Phi}}_{\mathbf{y}}(k,l)\mathbf{e}_{\{L,R\}}}, \tag{8}$$

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

$$\hat{\boldsymbol{\Phi}}_{\mathbf{y}}(k,l) = \alpha \boldsymbol{y}(k,l)\boldsymbol{y}^H(k,l) + (1-\alpha)\,\hat{\boldsymbol{\Phi}}_{\mathbf{y}}\,(k,l-1)\,, \tag{9}$$

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 [34], the corresponding DOA $\hat{\theta}_{\text{RTF}}(k,l)$ 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.,

$$\hat{\theta}_{\text{RTF}}(k,l) = \underset{\theta}{\text{argmin}}(|\hat{\kappa}(k,l) - \kappa(\theta)|), \tag{10}$$

with

$$\hat{\kappa}(k,l) = \frac{\mathbf{e}_L^T\hat{\boldsymbol{\Phi}}_{\mathbf{y}}(k,l)\mathbf{e}_R}{\left|\mathbf{e}_L^T\hat{\boldsymbol{\Phi}}_{\mathbf{y}}(k,l)\mathbf{e}_R\right|}, \tag{11}$$

$$\kappa(\theta) = \frac{\mathbf{e}_R^T\mathbf{a}_L(\theta)}{\left|\mathbf{e}_R^T\mathbf{a}_L(\theta)\right|}. \tag{12}$$

When the DOA $\hat{\theta}_{\text{RTF}}(k,l)$ is in a region of ±5° 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.,

$$\hat{\mathbf{a}}_{1,\{L,R\}}(k,l) = \beta\tilde{\mathbf{a}}_{\{L,R\}}(k,l) + (1-\beta)\,\hat{\mathbf{a}}_{1,\{L,R\}}(k,l-1). \tag{13}$$

with β recursive smoothing constant. When the DOA $\hat{\theta}_{\text{RTF}}(k,l)$ is in a region of ±5° 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.,

$$\hat{\mathbf{a}}_{2,\{L,R\}}(k,l) = \beta\tilde{\mathbf{a}}_{\{L,R\}}(k,l) + (1-\beta)\,\hat{\mathbf{a}}_{2,\{L,R\}}(k,l-1). \tag{14}$$

IV. Cognitive-Driven Binaural Beamformer

In this section, we present the proposed cognitive-driven speech enhancement system (see Fig. 1), consisting of three main blocks. Section IV-A describes the reference signal generation, where either MVDR or LCMV beamformers generate reference signals for AAD using the estimated DOA-based RTF vectors. Section IV-B reviews the least-squares-based AAD method in [3], 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 IV-C.

A. Reference Signal Generation Using Beamformers

In [10], [14] it has been shown that the decoding performance of the least-squares-based AAD method (see Section IV-B) 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 [19] it has been proposed to use an MVDR beamformer for generating reference signals. The MVDR beamformer [1], [35], [36] 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

$$\min_{\mathbf{w}} \underbrace{\mathbf{w}^H\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{w}}_{\text{noise output PSD}} \text{ subject to } \underbrace{\mathbf{w}^H\bar{\mathbf{a}}_t = 1}_{\text{target}}, \tag{15}$$

where $\boldsymbol{\Phi}_{\mathbf{v}} = \varepsilon\{\mathbf{v}\mathbf{v}^H\}$ 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 (15) is given by [1], [35], [36]

$$\mathbf{w}_{\text{MVDR}} = \frac{\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{a}}_t}{\bar{\mathbf{a}}_t^H\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{a}}_t}. \tag{16}$$

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 [35], [37], which adds an interference suppression constraint to the MVDR optimization problem in (15), i.e.,

$$\min_{\mathbf{w}} \underbrace{\mathbf{w}^H\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{w}}_{\text{noise output PSD}} \text{ subject to } \underbrace{\mathbf{w}^H\bar{\mathbf{a}}_t = 1}_{\text{target}},\ \underbrace{\mathbf{w}^H\bar{\mathbf{a}}_i = 0}_{\text{interference}}, \tag{17}$$

where $\bar{\mathbf{a}}_i$ denotes the RTF vector corresponding to the interfering speaker. The LCMV beamformer solving (17) is given by [37], [38]

$$\mathbf{w}_{\text{LCMV}} = \boldsymbol{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}\left(\bar{\mathbf{C}}^H\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}\right)^{-1}\mathbf{b}, \tag{18}$$

with

$$\bar{\mathbf{C}} = [\bar{\mathbf{a}}_t \quad \bar{\mathbf{a}}_i]\,, \quad \mathbf{b} = [1 \quad 0]^T\,. \tag{19}$$

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 (16) with target angle θ_t is given by

$$\mathbf{w}_{\text{MVDR}}\,(\theta_t) = \frac{\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}\,(\theta_t)}{\mathbf{a}^H\,(\theta_t)\,\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}\,(\theta_t)}. \tag{20}$$

The LCMV beamformer in (18) with target angle θ_t and interfering angle θ_i is given by

$$\mathbf{w}_{\text{LCMV}}\left(\theta_t,\theta_i\right)=\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}\left(\mathbf{C}^H\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}\right)^{-1}\mathbf{b}, \tag{21}$$

with

$$\mathbf{C}=\left[\mathbf{a}\left(\theta_t\right)\quad \mathbf{a}\left(\theta_i\right)\right],\quad \mathbf{b}=\left[1\quad 0\right]^T. \tag{22}$$

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. 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 $\hat{\theta}_t\leq 0^\circ$, 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 $\hat{\theta}_t>0^\circ$.

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

$$z_1=z_{t,1}+z_{i,1}+z_{v,1}, \tag{23}$$

$$z_2=z_{t,2}+z_{i,2}+z_{v,2}, \tag{24}$$

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 V), 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).

B. 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 [3], which consists of a training and a decoding step.

1) Decoding Step: The EEG recordings are first segmented into trials (see Section V-B 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 [3] to reconstruct an estimate of the attended speech envelope $\hat{e}_a[i]$ using a trained spatio-temporal filter, i.e.,

$$\hat{e}_a\left[i\right]=\mathbf{g}^T\mathbf{r}\left[i\right],\ i=1\dots I, \tag{25}$$

with

$$\mathbf{g}=\left[\mathbf{g}_1^T\,\mathbf{g}_2^T\dots\,\mathbf{g}_C^T\right]^T, \tag{26}$$

$$\mathbf{g}_c=\left[g_{c,0}\ g_{c,1}\dots\ g_{c,J-1}\right]^T, \tag{27}$$

$$\mathbf{r}\left[i\right]=\left[\mathbf{r}_1^T\left[i\right]\,\mathbf{r}_2^T\left[i\right]\ \dots\ \mathbf{r}_C^T\left[i\right]\right]^T, \tag{28}$$

$$\mathbf{r}_c\left[i\right]=\left[r_c\left[i+\Delta\right]\ r_c\left[i+1+\Delta\right]\ \dots\ r_c\left[i+J-1+\Delta\right]\right]^T, \tag{29}$$

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 (25) and the envelope of two reference signals are computed as

$$\rho_1=\rho\left(e_1\left[i\right],\hat{e}_a\left[i\right]\right),\ \rho_2=\rho\left(e_2\left[i\right],\hat{e}_a\left[i\right]\right), \tag{30}$$

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

$$\begin{cases}\hat{\theta}_a=\hat{\theta}_1,\quad \hat{\theta}_u=\hat{\theta}_2\quad \text{if}\quad \rho_1>\rho_2\\ \hat{\theta}_a=\hat{\theta}_2,\quad \hat{\theta}_u=\hat{\theta}_1\quad \text{otherwise.}\end{cases} \tag{31}$$

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 (31).

2) Training Step: Prior to the decoding step, the spatio-temporal filter $\mathbf{g}$ in (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 [3], [8], [9], [14], i.e.,

$$\min_{\mathbf{g}}\frac{1}{I}\sum_{i=1}^{I}\left(e_a\left[i\right]-\mathbf{g}^T\mathbf{r}\left[i\right]\right)^2+\beta\mathbf{g}^T\mathbf{D}\mathbf{g}, \tag{32}$$

with $\mathbf{D}$ denoting the derivative matrix [9] and β denoting a regularization parameter. The filter minimizing the regularized least-squares cost function in (32) is equal to

$$\mathbf{g} = (\mathbf{Q} + \beta\mathbf{D})^{-1}\,\mathbf{q}, \tag{33}$$

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

$$\mathbf{Q} = \frac{1}{I}\sum_{i=1}^{I}\left(\mathbf{r}\,[i]\,\mathbf{r}^T\,[i]\right),\ \ \mathbf{q} = \frac{1}{I}\sum_{i=1}^{I}\left(\mathbf{r}\,[i]\,e_a\,[i]\right). \tag{34}$$

C. Binaural Beamformer

The estimated DOAs of the attended and the unattended speaker in (31) are then used to steer a binaural beamformer (BBEAM in Fig. 1), generating binaural output signals. For binaural beamforming, we will consider either the binaural MVDR or LCMV beamformer.

The binaural MVDR beamformer [21] aims at minimizing the PSD of the output noise component while passing signals arriving from the estimated attended DOA $\hat{\theta}_a$. Similarly to (16), the binaural MVDR beamformer for the left and the right hearing aid using the estimated RTF vectors is given by

$$\mathbf{w}_{\text{BMVDR,L}} = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\hat{\mathbf{a}}_{a,L}}{\hat{\mathbf{a}}_{a,L}^H\mathbf{\Phi}_{\mathbf{v}}^{-1}\hat{\mathbf{a}}_{a,L}}, \tag{35}$$

$$\mathbf{w}_{\text{BMVDR,R}} = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\hat{\mathbf{a}}_{a,R}}{\hat{\mathbf{a}}_{a,R}^H\mathbf{\Phi}_{\mathbf{v}}^{-1}\hat{\mathbf{a}}_{a,R}}, \tag{36}$$

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

$$\mathbf{w}_{\text{BMVDR,L}}(\hat{\theta}_a) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}_L(\hat{\theta}_a)}{\mathbf{a}_L^H(\hat{\theta}_a)\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}_L(\hat{\theta}_a)}, \tag{37}$$

$$\mathbf{w}_{\text{BMVDR,R}}(\hat{\theta}_a) = \frac{\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}_R(\hat{\theta}_a)}{\mathbf{a}_R^H(\hat{\theta}_a)\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{a}_R(\hat{\theta}_a)}. \tag{38}$$

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 [39], [40], 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 (18), the binaural LCMV beamformer for the left and the right hearing aid using the estimated RTF vectors is given by

$$\mathbf{w}_{\text{BLCMV,L}} = \mathbf{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}_L\left(\bar{\mathbf{C}}_L^H\mathbf{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}_L\right)^{-1}\mathbf{b}_L, \tag{39}$$

$$\mathbf{w}_{\text{BLCMV,R}} = \mathbf{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}_R\left(\bar{\mathbf{C}}_R^H\mathbf{\Phi}_{\mathbf{v}}^{-1}\bar{\mathbf{C}}_R\right)^{-1}\mathbf{b}_R, \tag{40}$$

with

$$\bar{\mathbf{C}}_L = \left[\hat{\mathbf{a}}_{a,L}\quad \hat{\mathbf{a}}_{u,L}\right],\quad \mathbf{b}_L = \left[1\quad \delta_L\right]^T, \tag{41}$$

$$\bar{\mathbf{C}}_R = \left[\hat{\mathbf{a}}_{a,R}\quad \hat{\mathbf{a}}_{u,R}\right],\quad \mathbf{b}_R = \left[1\quad \delta_R\right]^T, \tag{42}$$

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 $0 \le \delta_L \le 1$ and $0 \le \delta_R \le 1$ 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

$$\mathbf{w}_{\text{BLCMV,L}}(\hat{\theta}_a, \hat{\theta}_u) = \mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}_L\left(\mathbf{C}_L^H\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}_L\right)^{-1}\mathbf{b}_L, \tag{43}$$

$$\mathbf{w}_{\text{BLCMV,R}}(\hat{\theta}_a, \hat{\theta}_u) = \mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}_R\left(\mathbf{C}_R^H\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{C}_R\right)^{-1}\mathbf{b}_R, \tag{44}$$

with

$$\mathbf{C}_L = \left[\mathbf{a}_L(\hat{\theta}_a)\quad \mathbf{a}_L(\hat{\theta}_u)\right], \tag{45}$$

$$\mathbf{C}_R = \left[\mathbf{a}_R(\hat{\theta}_a)\quad \mathbf{a}_R(\hat{\theta}_u)\right]. \tag{46}$$

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 [39], [40]. The output signals of the binaural LCMV beamformer for the left and the right hearing aid are equal to

$$z_L = \text{ISTFT}\left\{\mathbf{w}_{\text{BLCMV,L}}^H(\hat{\theta}_a, \hat{\theta}_u)\boldsymbol{y}\right\}, \tag{47}$$

$$z_R = \text{ISTFT}\left\{\mathbf{w}_{\text{BLCMV,R}}^H(\hat{\theta}_a, \hat{\theta}_u)\boldsymbol{y}\right\}. \tag{48}$$

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

$$z_L = z_{a,L} + z_{u,L} + z_{v,L}, \tag{49}$$

$$z_R = z_{a,R} + z_{u,R} + z_{v,R}, \tag{50}$$

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.

V. 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

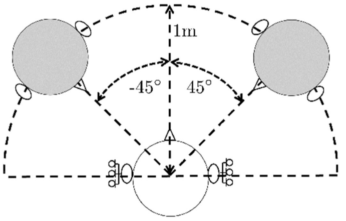


Fig. 2. Acoustic simulation setup for the reverberant condition. Two competing speakers were located at DOAs $\theta_1 = -45°$ and $\theta_2 = 45°$ and a distance of 1 m from the listener with two hearing aids, each equipped with 3 microphones.

estimation and beamforming algorithms, and the used performance measures.

A. 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 [28], and adding diffuse babble noise. The diffuse babble noise was simulated according to [41] using babble speech recordings and a cylindrically isotropic noise field assumption. The hearing aid setup in [28] 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 $\theta_1 = -45°$ and $\theta_2 = 45°$ (see Fig. 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.,

$$\mathrm{BSNR}_{in} = 10\log_{10}\frac{\phi_x}{\phi_v}, \tag{51}$$

with

$$\begin{aligned}\phi_x &= \varepsilon\left\{|x_{1,L,1}|^2\right\} + \varepsilon\left\{|x_{2,L,1}|^2\right\}\\ &\quad + \varepsilon\left\{|x_{1,R,1}|^2\right\} + \varepsilon\left\{|x_{2,R,1}|^2\right\},\\ \phi_v &= \varepsilon\left\{|v_{L,1}|^2\right\} + \varepsilon\left\{|v_{R,1}|^2\right\}.\end{aligned} \tag{52}$$

B. EEG Measurement and AAD Setup

Eighteen normal-hearing and German-speaking participants took part in this study (see [14]). 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 3 A).[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 [14] for more details). Similarly as in [3], [14], 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.

For filter training and decoding (see Section IV-B), 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 [8], followed by low-pass filtering at 8 Hz and downsampling 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 [14] for more details). All analyses were performed using the EEG recordings under the same experimental analysis condition. Similarly as in [19], the parameters of the spatio-temporal filter $\mathbf{g}$ in (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 I):

- 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.

[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.

TABLE I
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

C. *Algorithm Implementation Details*

1) DOA Estimation Algorithm: For the DOA estimation algorithm (see Section III-A), 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 [28] 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 (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 $\hat{\theta}_1 \leq \hat{\theta}_2$.

2) RTF Vector Estimation Algorithm: The RTF estimation algorithm (see Section III-B) was implemented using a weighted overlap-add (WOLA) framework with different STFT frame lengths, i.e., $\mathrm{FL} = 512,\ 1024,\ 2048,\ 4096,\ 8192$ samples, and an overlap of 50% between successive frames. The microphone covariance matrix in (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 (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 (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 (13) and (14) were recursively smoothed using a corresponding time constant of 100 ms.

As proposed in [42], 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 [28] 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 [28] for angle θ.

3) Beamforming: All considered beamformers were implemented using WOLA framework with a default STFT frame length $\mathrm{FL} = 512$ when using anechoic RTF vectors and a default STFT frame length $\mathrm{FL} = 8192$ 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 $\mathrm{FL} = 1024,\ 2048,\ 4096$.

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 IV-C), 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 (37) and (38).

D. *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 improvement (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

$$\mathrm{SINR}_{in,1} = 10\log_{10}\frac{\varepsilon\left\{|x_{1,L,1}|^2\right\}}{\varepsilon\left\{|x_{2,L,1}+v_{L,1}|^2\right\}}, \tag{53}$$

$$\mathrm{SINR}_{in,2} = 10\log_{10}\frac{\varepsilon\left\{|x_{2,R,1}|^2\right\}}{\varepsilon\left\{|x_{1,R,1}+v_{R,1}|^2\right\}}. \tag{54}$$

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

$$\mathrm{SINR}_{out,1} = 10\log_{10}\frac{\varepsilon\left\{|z_{t,1}|^2\right\}}{\varepsilon\left\{|z_{i,1}+z_{v,1}|^2\right\}}, \tag{55}$$

$$\mathrm{SINR}_{out,2} = 10\log_{10}\frac{\varepsilon\left\{|z_{t,2}|^2\right\}}{\varepsilon\left\{|z_{i,2}+z_{v,2}|^2\right\}}, \tag{56}$$

with all output signal components defined in (23) and (24). The average SINR improvement for both speakers is defined as

$$\Delta\mathrm{SINR} = \frac{\Delta\mathrm{SINR}_1 + \Delta\mathrm{SINR}_2}{2}, \tag{57}$$

with

$$\Delta\mathrm{SINR}_1 = \mathrm{SINR}_{out,1} - \mathrm{SINR}_{in,1}, \tag{58}$$

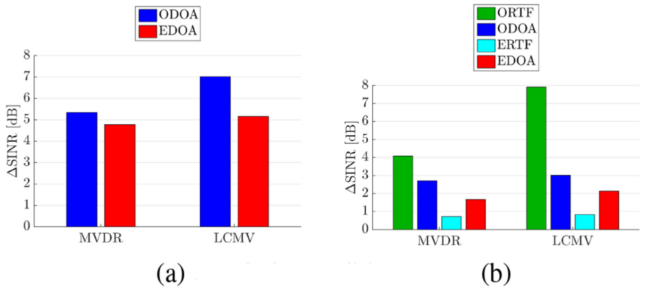
$$\Delta\mathrm{SINR}_2 = \mathrm{SINR}_{out,2} - \mathrm{SINR}_{in,2}. \tag{59}$$

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

$$\mathrm{BSINR}_{in} = 10\log_{10}\frac{\varepsilon\left\{|x_{a,L,1}|^2\right\} + \varepsilon\left\{|x_{a,R,1}|^2\right\}}{\varepsilon\left\{|x_{u,L,1}+v_{L,1}|^2\right\} + \varepsilon\left\{|x_{u,R,1}+v_{R,1}|^2\right\}}, \tag{60}$$

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



(a) (b)

Fig. 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.

$$\mathrm{BSINR}_{out} = 10\log_{10}\frac{\varepsilon\left\{|z_{a,L}|^2\right\} + \varepsilon\left\{|z_{a,R}|^2\right\}}{\varepsilon\left\{|z_{u,L}+z_{v,L}|^2\right\} + \varepsilon\left\{|z_{u,R}+z_{v,R}|^2\right\}}, \tag{61}$$

with all output signal components defined in (49) and (50). The BSINR improvement is defined as

$$\Delta\mathrm{BSINR} = \mathrm{BSINR}_{out} - \mathrm{BSINR}_{in}. \tag{62}$$

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 [43].

VI. 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 VI-A 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 VI-B we evaluate the decoding performance using these reference signals. Finally, in Section VI-C 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.

A. Performance of the Beamformers for Generating Reference Signals

For the anechoic and the reverberant condition, Fig. 3 depicts the average SINR improvement in (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

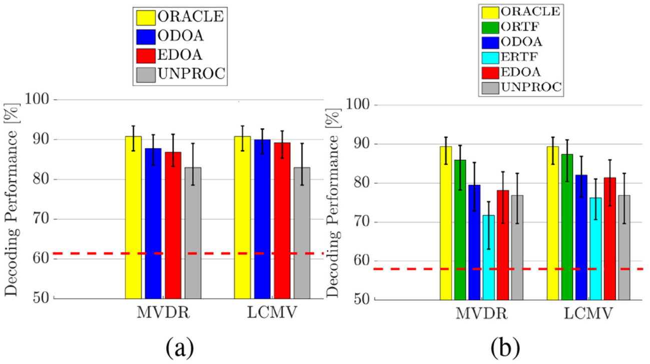


Fig. 4. Average decoding performance for (a) the anechoic condition and (b) the reverberant condition when using the oracle anechoic signals, the MVDR output signals, the LCMV output signals, and the unprocessed microphone signals as reference signals for decoding. The red dashed-line represents the 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.

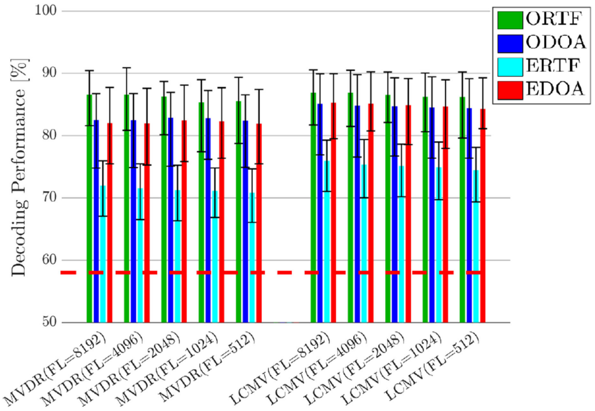


Fig. 5. Average decoding performance using different STFT frame lengths in the reverberant condition. The red dashed-line represents the upper boundary of the confidence interval corresponding to chance level based on a binomial test at the 5% significance level. The error bars represent the bootstrap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

interference suppression constraint in (17), which leads to a larger suppression of the interfering speaker (and a similar noise reduction) compared to the MVDR beamformers [35], [37]. 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 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 III-B 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).

B. Auditory Attention Decoding Performance

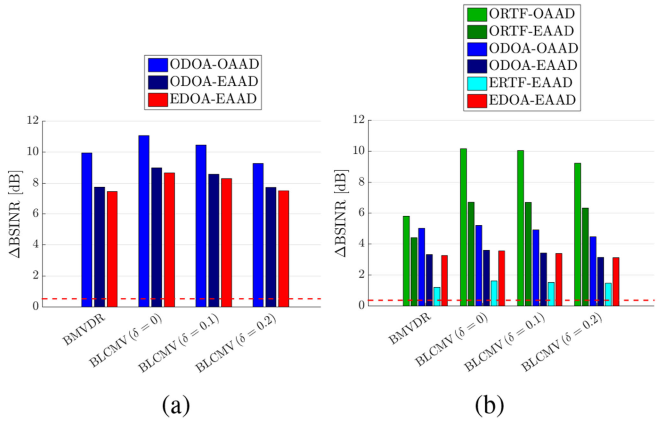
For the anechoic and the reverberant condition, Fig. 4 depicts the average decoding performance when using either the oracle anechoic signals, the MVDR output signals, the LCMV output signals, or the unprocessed microphone signals as reference signals for decoding. For both acoustic conditions, it can 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 (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). This is in accordance with the experimental results in [10], [14], 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. 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.

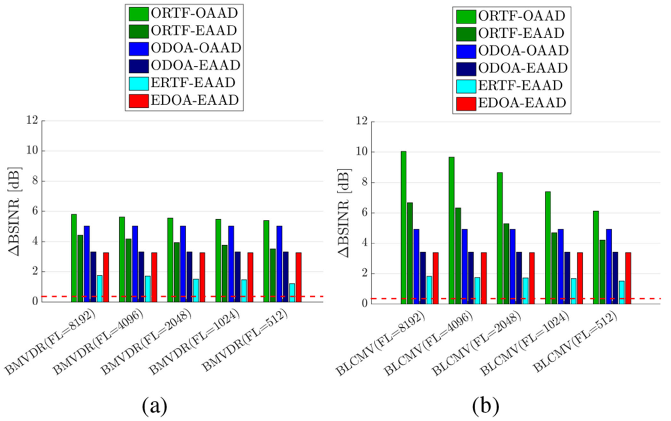
C. Binaural Speech Enhancement Performance

For the anechoic and the reverberant condition, Fig. 6 depicts the binaural SINR improvement of the complete proposed



(a) (b)

Fig. 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).



(a) (b)

Fig. 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).

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 vectors (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. 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 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 (31) using 30 second trials corresponds to 30 s. The latency caused by DOA estimation in (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 (37), (38), (43) and (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.

TABLE II
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

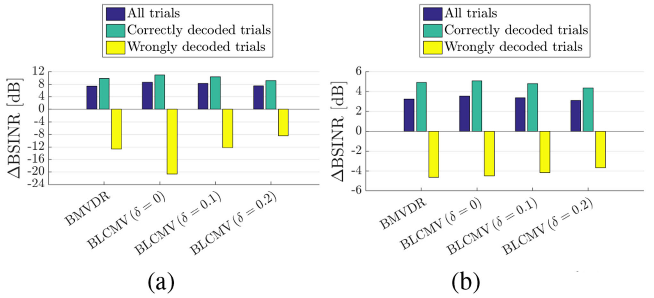


Fig. 8. Binaural SINR improvement averaged over all trials, all correctly decoded trials and all wrongly decoded trials obtained when using estimated DOAs and estimated AAD (EDOA–EAAD) for (a) the anechoic condition and (b) the reverberant condition.

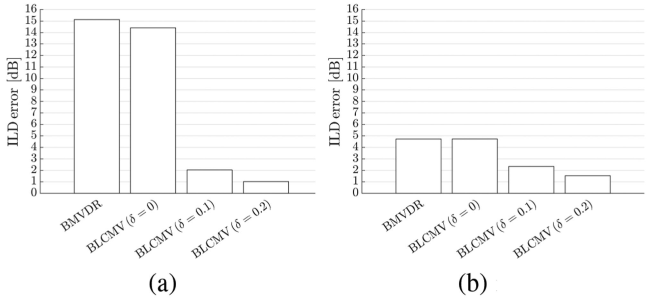


Fig. 9. ILD errors of the unattended speaker averaged over correctly decoded trials when using estimated DOAs and estimated AAD (EDOA–EAAD) for (a) the anechoic condition and (b) the reverberant condition.

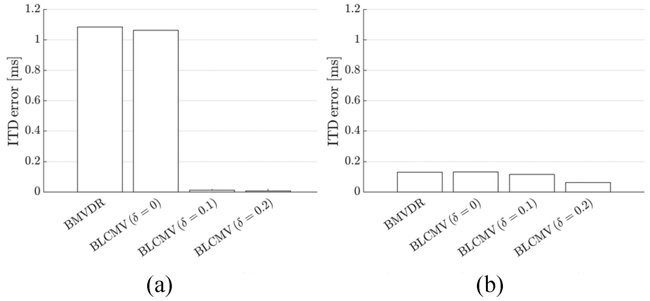


Fig. 10. ITD errors of the unattended speaker averaged over correctly decoded trials when using estimated DOAs and estimated AAD (EDOA–EAAD) for (a) the anechoic condition and (b) the reverberant condition.

To further investigate the impact of AAD errors on the binaural SINR improvement when using estimated AAD and estimated DOAs (EDOA–EAAD), Fig. 8 depicts the binaural SINR improvement averaged over all trials (as in Fig. 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 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, Figs. 9 and 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 IV-C). 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 [39], [40].

Table II 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.

VII. 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-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 [44] 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.

Acknowledgment

The authors would like to thank Nico Gößling for his fruitful discussions concerning the reverberant RTF vector estimation algorithm. They would also like to thank the anonymous reviewers for their comments that have improved the clarity of this paper.

References

[1] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[3] J. A. O'Sullivan *et al.*, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–706, 2014.

[4] C. Horton, R. Srinivasan, and M. D'Zmura, "Envelope responses in single-trial EEG indicate attended speaker in a cocktail party," *Neural Eng.*, vol. 11, no. 4, 2014, Art. no. 46015.

[5] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners attention from electroencephalography evoked by continuous speech," *Eur. J. Neurosci.*, pp. 1–8, 2017.

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

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

[8] 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 Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, May 2017.

[9] A. Aroudi and S. Doclo, "EEG-based auditory attention decoding using unprocessed binaural signals in reverberant and noisy conditions," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Jeju, South Korea, 2017, pp. 484–488.

[10] A. Aroudi and S. Doclo, "EEG-based auditory attention decoding: Impact of reverberation, noise and interference reduction," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Banff, Canada, Oct. 2017, pp. 3042–3047.

[11] 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.

[12] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: Boundary conditions for background noise and speaker positions," *J. Neural Eng.*, vol. 15, no. 6, 2018, Art. no. 66017.

[13] T. Dau, J. Maercher Roersted, S. Fuglsang, and J. Hjortkjær, "Towards cognitive control of hearing instruments using EEG measures of selective attention," *J. Acoust. Soc. Amer.*, vol. 143, no. 3, pp. 1744–1744, 2018.

[14] 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 Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 652–663, Apr. 2019.

[15] 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. [Online]. Available: https://doi.org/10.1101/218727

[16] J. O'Sullivan *et al.*, "Neural decoding of attentional selection in multi-speaker environments without access to clean sources," *J. Neural Eng.*, vol. 14, no. 5, 2017, Art. no. 56001.

[17] 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 Trans. Biomed. Eng.*, vol. 64, no. 5, pp. 1045–1056, May 2017.

[18] 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 *Proc. Eur. Signal Process. Conf.*, Kos, Greece, Aug. 2017, pp. 1660–1664.

[19] A. Aroudi, D. Marquardt, and S. Doclo, "EEG-based auditory attention decoding using steerable binaural superdirective beamformer," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Calgary, Canada, Apr. 2018, pp. 851–855.

[20] A. Aroudi and S. Doclo, "Cognitive-driven binaural LCMV beamformer using EEG-based auditory attention decoding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brighton, U.K., May 2019, pp. 406–410.

[21] B. Cornelis, S. Doclo, T. van den Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[22] 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 Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2016–2030, Sep. 2012.

[23] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 515–528, Mar. 2018.

[24] H. Kayser and J. Anemüller, "A discriminative learning approach to probabilistic acoustic source localization," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Juan-les-Pins, France, Sep. 2014, pp. 99–103.

[25] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[26] J. Chen, J. Benesty, and Y. A. Huang, "Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 1, pp. 25–36, Jan. 2005.

[27] M. Jeub, M. Drbecker, and P. Vary, "A semi-analytical model for the binaural coherence of noise fields," *IEEE Signal Process. Lett.*, vol. 18, no. 3, pp. 197–200, Mar. 2011.

[28] 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 J. Adv. Signal Process.*, vol. 2009, 2009, Art. no. 298605.

[29] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[30] 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 Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, p. 785–799, Apr. 2014.

[31] 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 Int. Conf. Acoust., Speech Signal Process.*, Brisbane, QLD, Australia, Apr. 2015, pp. 544–548.

[32] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1507–1519, Oct. 2019.

[33] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[34] D. Marquardt and S. Doclo, "Noise power spectral density estimation for binaural noise reduction exploiting direction of arrival estimates," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2017, pp. 234–238.

[35] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[36] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*. Berlin, Germany: Springer, 2001, pp. 19–38.

[37] E. Habets, J. Benesty, and P. A. Naylor, "A speech distortion and interference rejection constraint beamformer," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 854–867, Mar. 2012.

[38] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[39] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 543–558, Mar. 2016.

[40] 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. Int. Workshop Acoust. Signal Enhancement*, Tokyo, Japan, Sep. 2018, pp. 381–385.

[41] E. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[42] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Hands-free Speech Commun. Microphone Arrays*, Mar. 2017, pp. 11–15.

[43] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," vol. 53, no. 5, pp. 592–605, 2011.

[44] 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 Hearing*, vol. 22, pp. 1–12, Jan. 2018.

Ali Aroudi photograph and biography not available at the time of publication.

Simon Doclo (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007, he was a Postdoctoral Fellow with the Research Foundation Flanders, Electrical Engineering Department, Katholieke Universiteit Leuven, and the Cognitive Systems Laboratory, McMaster University, Canada. From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors, Leuven, Belgium. Since 2009, he has been a Full Professor with the University of Oldenburg, Oldenburg, Germany, and a Scientific Advisor for the Division Hearing, Speech and Audio Technology, Fraunhofer Institute for Digital Media Technology. His research activities center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks and hearing aid processing. He received several best paper awards, including International Workshop on Acoustic Echo and Noise Control 2001, EURASIP Signal Processing 2003, IEEE Signal Processing Society 2008, VDE Information Technology Society 2019. He is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing and the EAA Technical Committee on Audio Signal Processing. He was and is involved in several large-scale national and European research projects (ITN DREAMS, Cluster of Excellence Hearing4all, CRC Hearing Acoustics). He was a Technical Program Chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics in 2013 and a Chair of the ITG Conference on Speech Communication in 2018. In addition, he was a Guest Editor for several special issues, including IEEE Signal Processing Magazine and Elsevier's *Signal Processing* and was an Associate Editor for IEEE/ACM Transactions on Audio, Speech, and Language Processing and *EURASIP Journal on Advances in Signal Processing*.