2020 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2020, ESPOO, FINLAND

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

Ali Aroudi[†⋆] *Marc Delcroix*[†] *Tomohiro Nakatani*[†] *Keisuke Kinoshita*[†] *Shoko Araki*[†] *Simon Doclo*[⋆]

[†] NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
[⋆] Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, University of Oldenburg, Oldenburg, Germany

ABSTRACT

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. 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. The proposed system allows to enhance the attended speaker and jointly suppress reverberation, the interfering speaker and ambient noise. 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 reverberant and noisy conditions.

Index Terms— auditory attention decoding, convolutional beamformer, speech enhancement, mask estimation, EEG, dereverberation

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 [1, 2, 3, 4]. 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 [5, 6, 7, 8]. 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 [5], multi-channel Wiener filters [6] or minimum variance distortionless response (MVDR) beamformers [8]. 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 [7, 8].

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 [9]. 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.

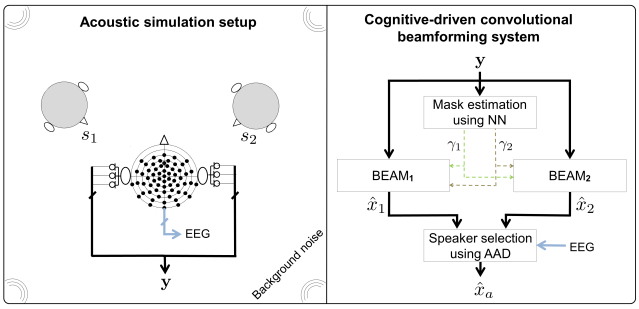

Fig. 1: Acoustic simulation setup and block diagram of the proposed cognitive-driven convolutional beamforming system.

The proposed system is depicted in Fig. 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 [10]. 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 direction-of-arrivals (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.

978-1-7281-6662-9/20/$31.00

2. COGNITIVE-DRIVEN CONVOLUTIONAL BEAMFORMER

2.1. Signal model

We consider an acoustic scenario comprising I competing speakers[1] with the clean signals denoted as $s_i[n]$, $i=1...I$ where n is the discrete time index. We consider a binaural hearing aid setup with M microphones. The m-th microphone signal $y_m[n]$ can be decomposed as

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

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 [9].

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

$$\mathbf{y}_{k,f}=[Y_{1,k,f}...Y_{M,k,f}]^T\in\mathbb{C}^{M\times 1}, \tag{2}$$

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

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 2.3).

Several neural network-based speech separation approaches have been proposed, both in frequency-domain and in time-domain [11, 12]. In this paper we use a BLSTM-based frequency-domain approach [11] since it trains faster than time-domain approaches such as [12], 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.,

$$[\mathbf{\Gamma}_{1,m}...\mathbf{\Gamma}_{I+1,m}]=h(\mathbf{Y}_m), \tag{3}$$

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 $i=1\ ...\ I$, 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) [11] with a scale-dependent SNR loss in the time-domain [13]. 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 (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.

[1]It should be noted that we provide a general description of the algorithms for I speakers, but limit our experiments in Section 4 to two speakers.

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. 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 3.

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 [1]. 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.,

$$\rho_i=\rho(\hat{e}_i[l],\ \hat{e}_a[l]),\ \ i=1...I, \tag{4}$$

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.,

$$\hat{x}_a[n]=\hat{x}_{\bar{i}}[n],\ \ \bar{i}=\underset{i}{\text{argmax}}\ \rho_i. \tag{5}$$

3. BEAMFORMING

In this section, we review the wMPDR convolutional beamformer [14], 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.

3.1. Weighted MPDR convolutional beamformer

The wMPDR convolutional beamformer in [14] 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

$$z_k=\bar{\mathbf{w}}^H\bar{\mathbf{y}}_k=\mathbf{w}_0^H\mathbf{y}_k+\sum_{\tau=b}^{L_w-1}\mathbf{w}_\tau^H\mathbf{y}_{k-\tau}, \tag{6}$$

where $\bar{\mathbf{w}}=\left[\mathbf{w}_0^T\ \mathbf{w}_b^T...\mathbf{w}_{L_w-1}^T\right]^T\in\mathbb{C}^{M(L_w-b+1)\times 1}$, $\bar{\mathbf{y}}_k=\left[\mathbf{y}_k^T\ \tilde{\mathbf{y}}_k^T\right]^T\in\mathbb{C}^{M(L_w-b+1)\times 1}$, $\tilde{\mathbf{y}}_k$ consists of the observation from b frames in the past until L_w-1 frames in the past, i.e., $\tilde{\mathbf{y}}_k=\left[\mathbf{y}_{k-b}^T...\mathbf{y}_{k-L_w+1}^T\right]^T$,

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 [10] 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 (6) can hence be written as dereverberation filtering followed by beamforming [10], i.e.,

$$\mathbf{d}_k = \underbrace{\mathbf{y}_k - \mathbf{G}^H \bar{\mathbf{y}}_k}_{\text{dereverberation}}, \quad z_k = \underbrace{\mathbf{q}^H \mathbf{d}_k}_{\text{beamforming}}. \tag{7}$$

Assuming that the output of the convolutional beamformer z_k follows a zero mean complex Gaussian distribution with a time-varying variance [14], the wMPDR convolutional 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.,

$$\mathcal{L}(\bar{\mathbf{w}}) \propto \frac{1}{K}\sum_{k=1}^{K}\left(-\ln(\lambda_k) - \frac{|z_k|^2}{\lambda_k}\right), \tag{8}$$

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 [14]

$$\max_{\bar{\mathbf{w}}} -\bar{\mathbf{w}}^H \bar{\mathbf{R}}_{\bar{\mathbf{y}}} \bar{\mathbf{w}} \quad \text{s.t.} \quad \underbrace{\bar{\mathbf{w}}^H \bar{\mathbf{a}} = 1}_{\text{target}}, \tag{9}$$

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 (9) is given by [10]

$$\bar{\mathbf{w}}_{\text{wMPDR}} = -\mathbf{G}\mathbf{q}_{\text{wMPDR}}, \tag{10}$$

where

$$\mathbf{G} = \mathbf{R}_{\tilde{y}}^{-1}\mathbf{P}_{\tilde{y}}, \quad \mathbf{q}_{\text{wMPDR}} = \frac{\mathbf{R}_d^{-1}\bar{\mathbf{a}}}{\bar{\mathbf{a}}^H \mathbf{R}_d^{-1}\bar{\mathbf{a}}}, \tag{11}$$

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 (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 [15], i.e.,

$$\bar{\mathbf{a}} = \mathbf{R}_{\tilde{t}+v}\text{MaxEig}\left(\mathbf{R}_{\tilde{t}+v}^{-1}\mathbf{R}_t\right), \tag{12}$$

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 $\mathbf{R}_{\tilde{t}+v} = \frac{\sum_k (1-\gamma_{t,k})\mathbf{d}_k\mathbf{d}_k^H}{\sum_k (1-\gamma_{t,k})}$ 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 (7) is performed using $\mathbf{G}$ in (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 (12) to steer the beamformer $\mathbf{q}_{\text{wMPDR}}$ in (11). Using the steered beamformer, the output signal z_k in (7) is obtained. The variance of the target speech component is then updated as $\lambda_k = |z_k|^2$ for the next iteration.

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.,

$$\max_{\bar{\mathbf{w}}} -\bar{\mathbf{w}}^H \bar{\mathbf{R}}_{\bar{\mathbf{y}}} \bar{\mathbf{w}} \quad \text{s.t.} \quad \underbrace{\bar{\mathbf{w}}^H \bar{\mathbf{a}} = 1}_{\text{target}}, \ \underbrace{\bar{\mathbf{w}}^H \bar{\mathbf{B}} = \boldsymbol{\delta}}_{\text{interference}}, \tag{13}$$

where $\bar{\mathbf{B}} = \left[\bar{\mathbf{b}}_1 \ldots \bar{\mathbf{b}}_U\right]$ contains the RETF vectors of U interfering speakers, with $U = I-1$, and $\boldsymbol{\delta} = [\delta_1 \ldots \delta_U]$ containing the interference suppression parameters, which control the amount of suppression of the interfering speakers. This optimization problem is the same as the optimization problem of the conventional LCMP beamformer [16], but with different relative transfer function (RTF) vectors and covariance matrix. Therefore the wLCMP convolutional beamformer can be obtained as

$$\bar{\mathbf{w}}_{\text{wLCMP}} = -\mathbf{G}\mathbf{q}_{\text{wLCMP}}, \tag{14}$$

where the dereverberation matrix $\mathbf{G}$ is obtained as in (11) and the beamforming vector $\mathbf{q}_{\text{wLCMP}}$ is obtained as in [16], i.e.,

$$\mathbf{q}_{\text{wLCMP}} = \mathbf{R}_d^{-1}\bar{\mathbf{C}}\left(\bar{\mathbf{C}}^H \mathbf{R}_d^{-1}\bar{\mathbf{C}}\right)^{-1}\mathbf{p}, \tag{15}$$

with $\bar{\mathbf{C}} = \left[\bar{\mathbf{a}} \ \ \bar{\mathbf{B}}\right]$ and $\mathbf{p} = \left[1 \ \ \boldsymbol{\delta}\right]^T$. Setting δ_u to zero in (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 (15) is estimated using (12). The RETF vector of the u-th interfering speaker $\bar{\mathbf{b}}_u$ is estimated as

$$\bar{\mathbf{b}}_u = \mathbf{R}_{\tilde{u}+v}\text{MaxEig}\left(\mathbf{R}_{\tilde{u}+v}^{-1}\mathbf{R}_u\right) \tag{16}$$

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 $\mathbf{R}_{\tilde{u}+v} = \frac{\sum_k (1-\gamma_{u,k})\mathbf{d}_k\mathbf{d}_k^H}{\sum_k (1-\gamma_{u,k})}$.

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

3.3. Relation with conventional MPDR and LCMP beamformers

The conventional MPDR beamformer aims at minimizing the power spectral density (PSD) of the output signal while preserving the reverberant target speech component in one of the microphone signals [17]. The MPDR beamformer is given by

$$\mathbf{w}_{\text{MPDR}} = \frac{\mathbf{R}_y^{-1}\mathbf{a}}{\mathbf{a}^H \mathbf{R}_y^{-1}\mathbf{a}}, \tag{17}$$

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 (17) is similar to the convolutional wMPDR beamformer in (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 LCMP beamformer incorporating interfering speaker suppression constraints and the wLCMP convolutional beamformer in (15). The conventional LCMP beamformer is given by [16]

$$\mathbf{w}_{\text{LCMP}} = \mathbf{R}_y^{-1}\mathbf{C}\left(\mathbf{C}^H \mathbf{R}_y^{-1}\mathbf{C}\right)^{-1}\mathbf{p}, \tag{18}$$

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

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

$$z_k = \mathbf{w}_{\{\text{MPDR, LCMP}\}}^H \mathbf{y}_k. \quad (19)$$

These output signals are obviously computed without involving a dereverberation step compared to the output signals of wMPDR and wLCMP convolutional beamformers in (6).

4. EXPERIMENTAL SETUP

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 [18], and adding diffuse babble noise, simulated according to [19]. The hearing aid setup in [18] 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.

4.2. 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 [20] 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.

4.3. Beamforming

All considered beamformers were implemented using a weighted overlap-add (WOLA) framework with an STFT frame length $\text{FL} = 512$, 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 (3).

We also compare our proposed system with a state-of-the-art cognitive-driven system proposed in [8], 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 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 parameters δ_1,...,δ_U were set to 0.1. Similarly as in [8], the DOAs of both speakers were estimated using a classification-based method [21] and the anechoic RTF vectors corresponding to the estimated DOAs were selected from a database of (measured) prototype RTF vectors [18].

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 [8] for details about the EEG recording and the AAD training and decoding configuration.

For the AAD training and decoding steps (see Section 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.

4.5. 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 Section2.4.

5. EXPERIMENTAL RESULTS

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

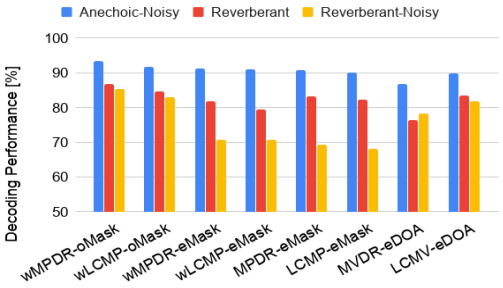

Fig. 2: Average auditory attention decoding performance for the anechoic-noisy, reverberant and reverberant-noisy conditions for the different considered beamformers. The upper boundary of the confidence interval corresponding to chance level for the anechoic-noisy, reverberant and reverberant-noisy conditions are 61.39%, 66.19%, 61.39%, respectively, computed based on a binomial test at the 5% significance level.

beamforming system. In Section 5.1 we investigate the impact of mask estimation errors on the AAD performance. In Section 5.2, we investigate the impact of AAD errors on the speech enhancement performance.

5.1. Auditory attention decoding performance

Figure 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.

5.2. Speech enhancement performance

Figure 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 3b depicts the fwSSNR improvement of the complete proposed system for the anechoic-noisy, reverberant and reverberant-noisy

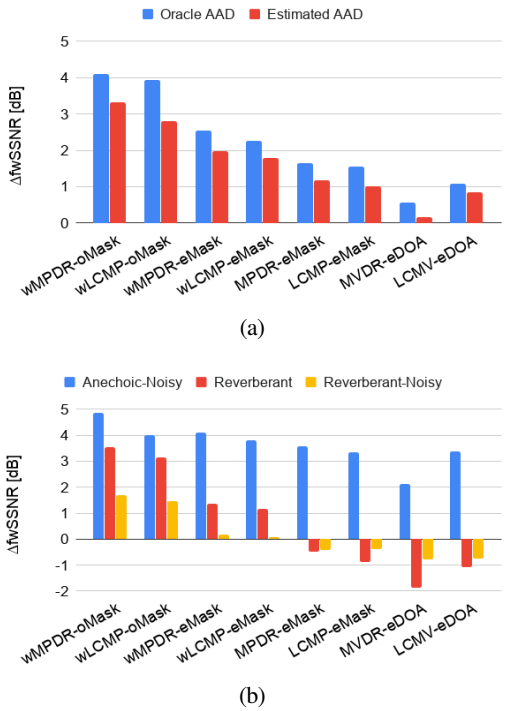

Fig. 3: fwSSNR improvement (a) averaged over all considered acoustic conditions when using oracle AAD and estimated AAD (b) for the anechoic-noisy, reverberant and reverberant-noisy conditions when using estimated AAD. The input fwSSNR averaged over all considered acoustic conditions is 2.06dB and the input fwSSNRs for the anechoic-noisy, reverberant and reverberant-noisy conditions are 2.9dB, 3.5dB, 0.5dB, respectively.

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.

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 beam-

former. 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.

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. We experimentally compared the proposed system with state-of-the-art cognitive-driven speech enhancement systems based on MVDR, LCMV, MPDR and LCMP beamformers. 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 the state-of-the-art cognitive-driven speech enhancement systems.

7. REFERENCES

[1] 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.

[2] 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.

[3] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Frontiers in Neuroscience*, vol. 13, p. 153, 2019.

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

[5] 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.

[6] 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.

[7] W. Pu, J. Xiao, T. Zhang, and Z. Luo, "A joint auditory attention decoding and adaptive binaural beamforming algorithm for hearing devices," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2019, pp. 311–315.

[8] 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.

[9] 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.

[10] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly optimal dereverberation and beamforming," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2020, pp. 216–220.

[11] 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.

[12] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2018, pp. 696–700.

[13] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 626–630.

[14] T. Nakatani and K. Kinoshita, "Maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation," in *Proc. of the European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, Sep. 2019, pp. 1–5.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 5206–5210.

[21] 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.