A JOINT AUDITORY ATTENTION DECODING AND ADAPTIVE BINAURAL BEAMFORMING ALGORITHM FOR HEARING DEVICES

Wenqiang Pu[1,3]*, Jinjun Xiao*[2]*, Tao Zhang*[2]*, Zhi-Quan Luo*[3]

[1] National Lab of Radar Signal Processing, Xidian University, Xi'an, China
[2] Starkey Hearing Technologies, Eden Prairie, MN 55344, USA
[3] Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

Traditional adaptive binaural beamforming algorithms for hearing devices often assume that the target talker is known or can be derived from the listener's look direction. When this assumption is violated, the traditional beamforming algorithms often produce distorted target speech and less than optimal noise and interference suppression. Recent advances in electroencephalography (EEG) and its applications to auditory attention decoding have offered a potential solution for tracking the listeners auditory attention in a multi-talker environment [2–5]. In this paper, we propose a unified model for joint auditory attention decoding and adaptive binaural beamforming, and solve the problem using an iterative optimization approach. The proposed algorithm has two advantages over the existing algorithms. First, the optimization objective aims to balance auditory attention alignment, target speech distortion, noise and interference suppression. Secondly, there is no need to estimate the speech envelope of each talker from the noisy and reverberant mixture which is a very challenging problem in practice. The proposed algorithm was evaluated using a newly recorded EEG database for a multi-talker, noisy and reverberant environment [6]. The evaluation results confirm the benefits of the proposed algorithm.

Index Terms— EEG signals, auditory attention, microphone array signal processing, acoustic beamforming

1. INTRODUCTION

In a multi-talker noisy and reverberant environment, humans have the unique capability to separate different sound sources and attend to a single source while ignoring other sound sources. In commonly available hearing devices, the listener's auditory attention is however unknown to the devices. And misalignment between the listener's attention and the target speaker can cause significant performance degradation of the speech enhancement algorithms in the devices. Understanding human auditory attention and improving target speech understanding in a multi-talker environment, such as in a cocktail party [7], has been an active research topic for decades [8–12]. Recent technology advances in electroencephalography (EEG) offers a potential non-invasive solution for tracking listener's auditory attention in such a environment. Using the signals collected from a scalp EEG system, various computational models [13–17] have been proposed to design a so-called *auditory attention decoder*, which attempts to reliably decode the auditory attention of a listener in a multi-talker environment.

Part of this wok is presented as a conference talk at ASA Spring 2018 in May this year [1].

In modern hearing aids, array signal processing algorithms [18–20] exploit the spatial diversity provided by the microphones for improved speech intelligibility and listening comfort. However, such algorithms including the multi-channel Wiener filter [21] and the linearly constrained minimum variance (LCMV) beamformer [22] often require the *a priori* knowledge of the auditory attention of listener. The multi-channel Wiener filter requires voice activity detection (VAD) for the attended talker and the LCMV beamformer needs to know which acoustic transfer function (ATF) corresponds to the attended talker. A common assumption that the target talker comes from the listener's look direction is not always true in practice.

Recent research has started incorporating the listener's auditory attention inferred from EEG signals into speech enhancement algorithms [2–4]. However, directly combining the auditory attention decoding results with beamforming algorithms has the following shortcomings. First, EEG signals usually have low signal to noise ratio which makes the attention decoder susceptible to decoding errors. Such errors can significantly degrade the speech enhancement algorithm performance. Secondly, the auditory attention decoder requires the source signal envelopes. In practice when such source signals are not available, an extra source separation processing is needed to extract the envelope of each source from the mixture, which itself is a challenging problem.

In this work, we propose a unified model for joint auditory attention decoding and binaural beamforming utilizing the tool of nonlinear optimization. The optimization objective aims for a balanced design to align auditory attention, control speech distortion, and reduce noise and interferences. Specifically, the attention is aligned by maximizing the Pearson correlation between the envelope of beamforming output and the linearly transformed EEG signal. The noise is reduced by minimizing the energy of the beamforming output subject to linear constraints suppressing the interferences and controlling the target speech distortion. Using a newly recorded EEG database for a multi-talker, noisy and reverberant environment, we compare the performance of the proposed algorithm with a baseline algorithm [4], which performs the attention decoding and beamforming at separate stages. The intelligibility-weighted signal to interference and noise ratio improvement (IW-SINRI) and spectral distortion (IW-SD) are used as performance metrics [21] to demonstrate the benefit of the proposed algorithm.

2. PROBLEM FORMULATION

Consider a listener equipped with a pair of binaural hearing aids with M microphones and L EEG electrodes. The listener attempts to listen to one target talker in the presence of $K-1$ competing talkers. The mic signals in the time-frequency domain can be expressed as

$$\boldsymbol{y}(\ell,\omega)=\sum_{k=1}^{K}\boldsymbol{h}_k(\omega)s_k(\ell,\omega)+\boldsymbol{n}(\ell,\omega), \tag{1}$$

where $\boldsymbol{y}(\ell,\omega)$ denotes the microphone signal at frame ℓ and frequency band ω ($\omega=1,2,\cdots,\Omega$); $\boldsymbol{h}_k(\omega)$ is the ATF [18] of the k-th speech source and $s_k(\ell,\omega)$ is the corresponding speech signal; and $\boldsymbol{n}(\ell,\omega)$ is the background noise in time-frequency domain. Further, $e_i(t)$ is used to denote the recorded EEG signals of EEG electrode i ($i=1,2,\cdots,L$) at time instance t.

We consider the binaural beamforming problem of linearly combining signals received at the M microphones using the attention information from the EEG signals. Specifically, consider a time period where the beamformer does not change and let $\boldsymbol{w}(\omega)$ be the beamformer coefficients at frequency band ω, then the output signals are $z(\ell,\omega)=\boldsymbol{w}^H(\omega)\boldsymbol{y}(\ell,\omega)$. The EEG signals $e_i(t)$ are utilized to direct the beamformer to point to the attended talker (see Fig. 1).

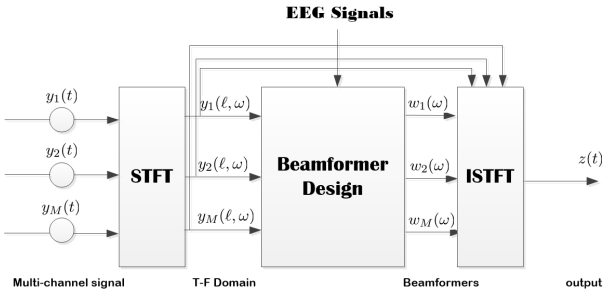

Fig. 1. Illustration of the proposed beamforming system.

We first give a brief description of a linear auditory attention decoder proposed in the literature [13, 16]. Let $s_a(t)$ denote the corresponding envelope of attended speech signal. A linear regression model is used to learn a linear mapping that reconstructs $s_a(t)$ from EEG signals $\{e_i(t)\}_{i=1}^{L}$. Specifically, let $\{g_i^*(\tau)\}$ be the learned linear mapping coefficients, then $s_a(t)$ is reconstructed as $\hat{s}_a(t)=\sum_{i=1}^{L}\sum_{\tau=\tau_1}^{\tau_2}g_i^*(\tau)e_i(t+\tau)$, where τ_1,τ_2 are specified time delays. The decoder is then simply to compare the Pearson correlation between $\hat{s}_a(t)$ and the envelope of each source. The one with the largest Pearson correlation is decoded as the attended source.

As a baseline algorithm for performance comparison purpose in Section 3, we propose to use auditory attention decoding followed by a LCMV beamformer, or AAD-LCMV for short [4]. In the source separation stage, this algorithm applies two different LCMV beamformers, i.e., use linear constraints to preserve one source and reject another, to separate the two source signals. The decoding is then done by comparing the Pearson correlations of the separated signals with respect to the constructed $\hat{s}_a(t)$. The decoded attention is finally used to perform a final LCMV beamforming the produce the output signal.

2.1. Joint Attention Decoding and Beamforming

Ideally, the beamformer enhances the attended source and suppresses the noise and interference. In such a case, the envelope of beamformer output can be used to replace the envelope of the attended source: $s_a(t)$ for the purpose of Pearson correlation calculation. This leads to one of the criteria for designing the beamformer by maximizing the Pearson correlation between the transformed EEG signal $\hat{s}_a(t)$ and the envelope of beamforming output.

To do so, we first give a mathematical description of calculating the envelope of beamforming output in terms of the microphone signals and beamforming weights. We assume the beamforming output of frame ℓ corresponding to time sample indexes $t,t+1,\cdots,t+N-1$, where N is the number of samples per STFT frame. Let $z(t),z(t+1),\cdots,z(t+N-1)$ denote the beamforming outputs in time domain with respect to frame ℓ. The beamforming output at this frame envelope are actually the absolute values of the corresponding analytic signal $\tilde{z}(t),\tilde{z}(t+1),\cdots,\tilde{z}(t+N-1)$, which can be represented by discrete Fourier transformation (DFT) [23] in terms of beamforming weights $\{\boldsymbol{w}(\omega)\}$ as

$$\begin{bmatrix}\tilde{z}(t)\\ \tilde{z}(t+1)\\ \vdots\\ \tilde{z}(t+N-1)\end{bmatrix}=\boldsymbol{D}_W\boldsymbol{F}\boldsymbol{D}_H\begin{bmatrix}\boldsymbol{w}(1)^H\boldsymbol{y}(\ell,1)\\ \boldsymbol{w}(2)^H\boldsymbol{y}(\ell,2)\\ \vdots\\ \boldsymbol{w}(\Omega)^H\boldsymbol{y}(\ell,\Omega).\end{bmatrix} \tag{2}$$

In (2), $\boldsymbol{D}_H\in\mathbb{R}^{\Omega\times\Omega}$ is a diagonal matrix for forming one-sided analytic signal [23], $\boldsymbol{F}\in\mathbb{C}^{\Omega\times\Omega}$ is the inverse of DFT matrix, and $\boldsymbol{D}_W\in\mathbb{R}^{\Omega\times\Omega}$ is a diagonal matrix for compensating the synthesis window used in STFT. For notational simplicity, (2) can be compactly expressed as

$$\tilde{z}(t+n)=\boldsymbol{w}^H\boldsymbol{u}_{\ell,n},\ n=0,1,\cdots,N-1, \tag{3}$$

where $\boldsymbol{w}=[\boldsymbol{w}(1)^H,\boldsymbol{w}(2)^H,\cdots,\boldsymbol{w}(\Omega)^H]^H\in\mathbb{C}^{M\Omega}$ is the concatenated beamforming vector and $\boldsymbol{u}_{\ell,n}\in\mathbb{C}^{M\Omega}$ is determined by $\{\boldsymbol{y}(\ell,\omega)\}$ and coefficients in matrix $\boldsymbol{D}_W,\boldsymbol{F}$, and $\boldsymbol{D}_H$. Based on (2) and (3), the envelopes of beamforming output of frame l are represented as $|\tilde{z}(t+n)|=|\boldsymbol{w}^H\boldsymbol{u}_{\ell,n}|,\ n=0,1,\cdots,N-1$.

Finally the Pearson correlation between $\hat{s}_a(t)$ and $|\tilde{z}(t)|$ is a function of beamformer, denoted as $\kappa(\{\boldsymbol{w}(\omega)\})$:

$$\kappa(\{\boldsymbol{w}(\omega)\})=\frac{\sum_{t=t_1}^{t_2}\bar{s}_a(t)\bar{z}(t)}{\sigma_s\sigma_z}, \tag{4}$$

where $\bar{s}_a(t)=\hat{s}_a(t)-\Lambda(\hat{s}_a(t'))$, $\bar{z}(t)=|\tilde{z}(t)|-\Lambda[|\tilde{z}(t')|]$, $\sigma_s=\sqrt{\Lambda[\bar{s}_a^2(t)]}$, $\sigma_z=\sqrt{\Lambda[\bar{z}^2(t)]}$. Here $\Lambda[x(t)]\triangleq\frac{1}{t_2-t_1+1}\sum_{t=t_1}^{t_2}x(t)$ is the average operation. Notice (4) assumes $\bar{s}_a(t)$ and $|\bar{z}(t)|$ are synchronized with a same sampling rate which is not true in practice. Hence, an extra synchronization procedure is needed but does not change the mathematical formula of $\kappa(\{\boldsymbol{w}(\omega)\})$, i.e., down sampling $\tilde{z}(t)$ from (3) according to the sampling rates of the audio and EEG signals.

Maximizing $\kappa(\{\boldsymbol{w}(\omega)\})$ guides the beamformer to focus on the attended source based on the auditory attention information in the EEG signals. Besides the protection of attended source, another beamformer design criterion is noise reduction. These two design criteria can be formulated as a combined objective function in the optimization problem and can be presented as follows:

(1) *Attention assignment maximizing Pearson correlation:* The first criterion of beamformer design is to assign the attention to different sources based on the information from the EEG signals. By exploiting the a priori information of the ATFs $\{\boldsymbol{h}_k(\omega)\}$, we propose to enforce a set of equality constraints, $\boldsymbol{w}(\omega)^H\boldsymbol{h}_k(\omega)=\alpha_k,\forall\omega$. This leads to the following problem

$$\max_{\{\boldsymbol{w}(\omega)\},\boldsymbol{\alpha}}\ \kappa(\{\boldsymbol{w}(\omega)\})\quad\text{s.t. }\boldsymbol{w}(\omega)^H\boldsymbol{h}_k(\omega)=\alpha_k,\ \forall k,\omega, \tag{5a}$$

$$\mathbf{1}^T\boldsymbol{\alpha}=1,\ \alpha_k\geq 0,\ \forall k. \tag{5b}$$

In (5), $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_K]^T$ are the weights of combining the K source signals in the beamforming output. And we refer it as the listener's auditory attention assignment among the K sources.

(2) *Noise energy reduction:* The energy of the background noise is reduced in minimum variance sense as

$$\min_{\boldsymbol{w}(\omega)} \mathbb{E}\left[|\boldsymbol{w}(\omega)^H \boldsymbol{n}(\omega)|^2\right] \equiv \min_{\boldsymbol{w}(\omega)} \boldsymbol{w}(\omega)^H \boldsymbol{R}_n(\omega)\boldsymbol{w}(\omega), \quad (6)$$

where $\boldsymbol{R}_n(\omega) \triangleq \mathbb{E}[\boldsymbol{n}(\omega)\boldsymbol{n}(\omega)]$ is the auto-correlation matrix of background noise.

Combining (6) into the objective function in (5), we obtain the proposed EEG-assisted beamforming formulation as follows:

$$\min_{\{\boldsymbol{w}(\omega)\},\boldsymbol{\alpha}} \sum_{\omega=1}^{\Omega} \underbrace{\boldsymbol{w}(\omega)^H \boldsymbol{R}_n(\omega)\boldsymbol{w}(\omega)}_{\text{noise reduction}} \underbrace{-\mu\kappa(\{\boldsymbol{w}(\omega)\})}_{\text{attention assignment}} \underbrace{-\gamma\|\boldsymbol{\alpha}\|^2}_{\text{sparsity}} \quad (7)$$

s.t. (5a), (5b).

In (7), we add an extra sparsity regularization term $-\gamma\|\boldsymbol{\alpha}\|^2$ on the attention assignment among sources. The nonnegative parameters μ and γ are pre-determined and used to obtain a desired tradeoff between noise reduction and attention assignment based on EEG signals in the beamformer design.

2.2. Optimization Algorithm

Problem (7) is a noncovex problem due to the nonlinear functions $-\mu\kappa(\{\boldsymbol{w}(\omega)\})$ and $-\gamma\|\boldsymbol{\alpha}\|^2$. As its constraints have a favorable form, i.e., separable with respect to $\{\boldsymbol{w}(\omega)\}$ across frequency bands ω in (5a), we adopt the well-known gradient projection method (GPM) [24] to solve Problem (7). The solver can be efficiently implemented by exploiting the separable property. With a proper stepsize rule (i.e., we use Armijo rule [24] in the experiment), the GPM solves problem (7) with guaranteed convergence (Proposition 2.3 in [24]). Since the major computation effort in the GPM for problem (7) is a projection onto the polyhedron defined by (5a) and (5b), we use the alternating direction method of multiplier (ADMM) [25] to solve the projection subproblem, which has a parallel implementation for primal updates with respect to $\{\boldsymbol{w}(\omega)\}$ in closed-forms. In short, the dominate computation complexity is $\mathcal{O}(\Omega(M^3 + KM^2))$ per GPM iteration. As the GPM and ADMM are standard algorithms, detailed derivation of solving Problem (7) is omitted due to space limitation.

2.3. Adaptive Beamforming Implementation

In an adaptive beamformer formulation, the beamformer $\{\boldsymbol{w}(\omega)\}$ is updated based on the new noise estimate and EEG signals in a new frame. In this subsection, we provide an adaptive updating scheme for the beamforming formulation in (7).

Suppose at time t, the latest EEG signals $\{e_i(t')\}, t' = t - T, t - T + 1, \ldots, t$ are used to update the Pearson correlation function $\kappa(\{\boldsymbol{w}(\omega)\})$. For the convenience of presentation, we denote $\kappa(\{\boldsymbol{w}(\omega)\})$ at time t as $\kappa_t(\{\boldsymbol{w}(\omega)\})$. Further, the noise correlation matrix at time t is denoted as $\boldsymbol{R}_{n,t}(\omega)$, and the objective function in (7) at time t is denoted by $f_t(\{\boldsymbol{w}(\omega)\}, \boldsymbol{\alpha})$, i.e., $f_t(\{\boldsymbol{w}(\omega)\}, \boldsymbol{\alpha}) \triangleq \sum_{\omega=1}^{\Omega} \boldsymbol{w}(\omega)^H \boldsymbol{R}_{n,t}(\omega)\boldsymbol{w}(\omega) - \mu\kappa_t(\{\boldsymbol{w}(\omega)\}) - \gamma\|\boldsymbol{\alpha}\|^2$.

Let us further introduce notation $\boldsymbol{x} \triangleq (\{\boldsymbol{w}(\omega)\}, \boldsymbol{\alpha})$. By the discussion in Section 2.2, the optimal $\boldsymbol{x}$ from minimizing $f_t(\boldsymbol{x})$ can be solved by GPM. At iteration r, the GPM updates are

$$\bar{\boldsymbol{x}} = [\boldsymbol{x}^r - s\nabla f_t^r]^+ \quad (8a)$$

$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r + \lambda^r(\bar{\boldsymbol{x}} - \boldsymbol{x}^r), \quad (8b)$$

where ∇f_t^r is the gradient of function $f_t(\boldsymbol{x})$ at $\boldsymbol{x}^r$, $[\cdot]^+$ denotes the projection operation onto the constraint set (5b) and (5a), s is a pre-determined positive scalar, and λ^r is the stepsize determined by the Armijo rule. In addition, due to the limited computation ability of hearing aids, it is necessary to reduce the computation cost of the adaptive implementation of the GPM updates (8) in an efficient manner. Hence we propose the following adaptive updating scheme (Algorithm 1) based on (8), where the solution of previous time window serves as the initial point for the current time window updates. This implementation is described in Table 1.

Algorithm 1 GPM Based Adaptive Updating Scheme

```
for time t = 0, 1, ..., do
    Update objective function f_t(x);
    Compute the gradient of f_t(x);
    Specify initial point x^0 = ({w_{t-1}(ω)}, α_{t-1})
    Fixed number of GPM updates (8);
    Update ({w_t(ω)}, α_t).
end for
```

3. EVALUATION

3.1. EEG Database

We collected EEG data from 12 normal-hearing subjects with signed consent in a multi-talker, noisy and reverberant environment. A set of binaural audio stimuli were generated using a set of ATFs for a simulated noisy and reverberant room. A room of size 8m×10m with height 3.6m is used in the simulation. The reverberation time is set to be 0.6 second. The hearing aids wearer is located at the center of the room. Each hearing aid has 2 microphones with 7.5mm spacing. There are 2 sources: one on the left and one on the right, with both being 1m away from the subject. The background babble noise was generated using sixteen loudspeakers distributed equally on the circle 2 meters away from the subject. The talkers were set to the same level and the babble noise is 5dB lower than the voice of each talker. The audio stimuli were presented to the subject using a set of ER-3 insert earphones at a loud but comfortable level in a sound-treated and semi-electrically shielded sound booth.

Each subject was instructed to perform a binaural listening task and attend to one of the talkers at a time. Each subject listened to stories with total duration of 20 minutes. Each story was split into multiple 1-minute segments. Each segment was chosen in a logical stopping point in the context of the story. Therefore no story segment ended in the middle of a word or phrase. The subjects were told to attend either the left talker or the right talker for the whole duration of a segment, and were not allowed to switch their attention within a segment. Thus, each segment had two stimuli speech tracks, one that will be referred to as the attended stimuli, and the other being the unattended stimuli. A 63-channel scalp EEG system was used to record the subjects response at a sampling frequency of 10 kHz. All recording was done on BrainVison Products hardware.

3.2. Evaluation Setup

In this section, we evaluate the performance of the proposed algorithm on the EEG database described in Section 3.1. The intelligibility-weighted signal-to-interference and noise ratio improvement (IW-SINRI) and intelligibility-weighted spectral distortion (IW-SD) are used as performance metrics [21].

In the evaluation, the EEG signals are down-sampled to 20 Hz. During the training stage for $\{g_i(\tau)\}$, all $L = 63$ EEG channels are used in the regression model and the latency of EEG signals is specified as $0 \sim 250$ ms, which corresponds to $\tau = 0, 1, \cdots, 5$ for 20 Hz sample rate. The regularization parameter for the regression model is fixed to be 5×10^{-3} [16]. The leave-one-out cross validation among the segments is used to train the coefficients $\{g_i^*(\tau)\}$. In the beamformer, a 512-point FFT with 50% overlap is used in STFT (Hanning window). Both the proposed and AAD-LCMV approaches utilize the anechoic relative transfer functions in the equality constraints and the correlation matrix $\{\boldsymbol{R}_n(\omega)\}$ is estimated by

sample averaging from a 5 seconds noise-only time period. In the EEG database presented in Section 3.1, the ground-truth direction of the attended source includes both the left and the right. We always compare the beamforming output to the front microphone on the far-side of the ground-truth. Due to the space limitation, the result of comparing to front microphone of the near-side is omitted.

3.3. Evaluation Results

In this section, we present the evaluation results. We choose two subjects only for illustration purpose since the algorithm performs similarly on all subjects. These two subjects are purposely chosen to have different levels of noise in the EEG signals: with 30 second segments, the decoding accuracy of the first subject is 93.02% and the second subject is 76.08%.

We first study the behavior of the proposed algorithm with off-line implementation, where we divide those one-minute records into 30s segments and each segment is used for evaluation. We set $\mu = 100$ and $\gamma = 0, 10, 100$. The IW-SINRI and IW-SD versus Pearson correlation difference $\Delta\rho = \rho_{\text{att}} - \rho_{\text{unatt}}$ is plotted in Fig. 2. One interesting observation is the continuous output of the IW-SINRI and IW-SD values in the proposed algorithm versus the bimodal distribution of that from the AAD-LCMV algorithm. In the proposed algorithm, the parameter γ can be increased to increase influence of the sparsity constraints and leads to a pattern that is similar to the AAD-LCMV algorithm.

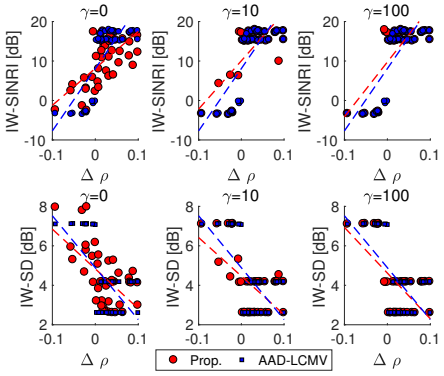

Fig. 2. IW-SINRI and IW-SD distributions (Sub. 2).

Then, we study the behavior of proposed algorithm in real time with an adaptive implementation. Each one-minute recording is selected as one trial for evaluation. EEG signals within a time window of length 10 seconds is used to update the beamformer. The time window is shifted every 2 seconds and 5 iterations of the GP are used for updating $\{\boldsymbol{w}(\omega)\}$ and $\boldsymbol{\alpha}$. The parameters μ and γ are specified as $\mu = 100$ and $\gamma = 0$, and the initial value of $\boldsymbol{\alpha}$ is set as $\alpha_1 = \alpha_2 = 0.5$.

The average IW-SINRI and IW-SD for all subjects over time are plotted in Fig. 3. The vertical lines correspond to one standard deviation of the 1-minute segments. Because the shortest segment is 54s and the algorithm needs 10s to generate result, the figures only show data between 10s∼54s. It can be observed that the proposed algorithm has similar IW-SINRI in some places and better in other places and slightly improved IW-SD than the AAD-LCMV algorithm, but it achieves smaller variation among segments due to the joint formulation.

To further understand the advantage of the proposed algorithm, we plot one representative segment for each subject in Fig. 4. In the cases where the decoding error happens at a sparse rate, the proposed joint algorithm enhances the attended talker consistently while

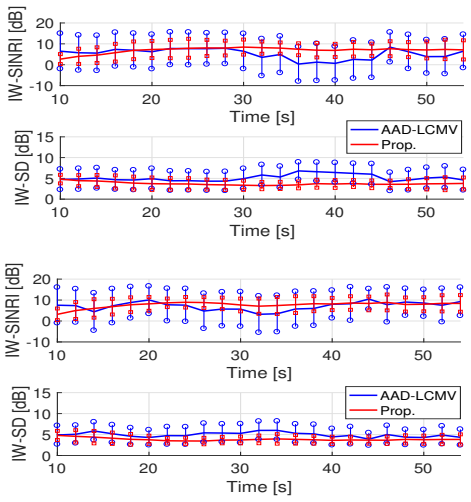

Fig. 3. Ave. IW-SINRI and IW-SD (Top: Sub. 1; Bottom: Sub. 2).

the AAD-LCMV is very susceptible to noise in the EEG signals and produces large errors in IW-SINRI or IW-SD when the attention decoding is inaccurate.

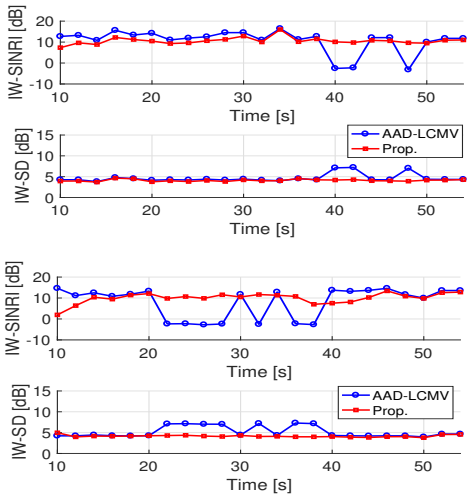

Fig. 4. IW-SINRI and IW-SD of representative segments (Top: Sub. 1; Bottom: Sub. 2).

4. CONCLUSION

This paper proposes a unified optimization model for joint auditory attention decoding and adaptive binaural beamforming, with the aim of balancing auditory attention alignment, target protection, and noise suppression. Furthermore, the proposed algorithm does not need to estimate the speech envelope of each talker from the noisy and reverberant mixture. A gradient projection based algorithm is proposed to efficiently solve the proposed optimization problem. The evaluation on a recorded EEG database in a multi-talker, noisy and reverberant environment demonstrates benefit of the proposed algorithm. As part of the future work, we plan to extend our algorithm evaluation to more realistic situations include attention switching. We also plan to explore the listener's response and adaptation to such an algorithm in a close-loop setup when enhanced speech signals are presented to the listener in real-time.

References

[1] W. Pu, J. Xiao, T. Zhang, and Z.-Q. Luo, "An optimization model for electroencephalography-assisted binaural beamforming," *The Journal of the Acoustical Society of America*, vol. 143, pp. 1744, 2018.

[2] J. O'Sullivan, Z. Chen, J. Herrero, G. McKhann, S. A. Sheth, A. D. Mehta, and N. Mesgarani, "Neural decoding of attentional selection in multi-speaker environments without access to clean sources.," *Journal of neural engineering*, vol. 14 5, pp. 056001, 2017.

[3] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: boundary conditions for background noise and speaker positions," *Journal of Neural Engineering*, vol. 15, no. 6, pp. 066017, 2018.

[4] A. Aroudi, D. Marquardt, and S. Doclo, "Cognitive-driven binaural speech enhancement system for hearing aid applications," *International Hearing Aid Research Conference*, Lake Tahoe, CA, 2018.

[5] S. Miran, S. Akram, A. Sheikhattar, J. Simon, T. Zhang, and B. Babadi, "Real-time tracking of selective auditory attention from m/eeg: A bayesian filtering approach," *Frontiers in neuroscience*, vol. 12, 2018.

[6] J. Xiao, S. Miran, B. Babadi, and T. Zhang, "An EEG database for a multi-talker, noisy and reverberant environment for hearing device applications," *International Hearing Aid Research Conferences (IHCON)*, Lake Tahoe, CA, 2018.

[7] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[8] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS biology*, vol. 10, no. 1, pp. e1001251, 2012.

[9] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[10] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 833–843, 2004.

[11] A. Cheveigne, D. Wong, G. Liberto, J. Hjortkjaer, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206 – 216, 2018.

[12] S. A. Fuglsang, T. Dau, and J. Hjortkjaer, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435 – 444, 2017.

[13] 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*, vol. 25, no. 7, pp. 1697–1706, 2015.

[14] A. J. Power, J. J. Foxe, E. Forde, R. B. Reilly, and E. C. Lalor, "At what time is the cocktail party? a late locus of selective attention to natural speech," *European Journal of Neuroscience*, vol. 35, no. 9, pp. 1497–1503, 2012.

[15] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications," *Journal of neural engineering*, vol. 12, no. 4, pp. 046007, 2015.

[16] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, 2017.

[17] J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay, *The auditory system at the cocktail party*, vol. 60, Springer, 2017.

[18] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2008.

[19] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, March 2015.

[20] W. Pu, J. Xiao, T. Zhang, and Z. Luo, "A penalized inequality-constrained minimum variance beamformer with applications in hearing aids," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2017, pp. 175–179.

[21] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of multichannel wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 487–503, July 2005.

[22] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 543–558, March 2016.

[23] L. Marple, "Computing the discrete-time analytic signal via FFT," *IEEE Transactions on Signal Processing*, vol. 47, no. 9, pp. 2600–2603, Sep 1999.

[24] D. P. Bertsekas, *Nonlinear programming*, Athena scientific Belmont, 1999.

[25] M. Hong and Z.-Q. Luo, "On the linear convergence of the alternating direction method of multipliers," *Mathematical Programming*, vol. 162, no. 1, pp. 165–199, 2017.