A CASCADED SEMI-BLIND SOURCE SEPARATION METHOD FOR JOINT ACOUSTIC ECHO CANCELLATION, INTERFERENCE SUPPRESSION, AND NOISE REDUCTION

Xianrui Wang[1,2], Kaien Mo[2], Yichen Yang[1,2], Liyuan Zhang[2], Shoji Makino[2], and Jingdong Chen[1]

[1]CIAIC and Shaanxi Provincial Key Laboratory of Artificial Intelligence,
Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[2]Waseda University, Kitakyushu, Japan

ABSTRACT

Acoustic echo cancellation (AEC), interference suppression, and noise reduction play important roles in full-duplex communication. However, conventional systems that cascade adaptive filters and beamformers often experience a significant degradation in performance during doubletalk situations. To tackle this issue, this paper presents a multichannel semi-blind-source-separation (SBSS) method that combines the element-wise iterative source steering (EISS) AEC algorithm with a geometrically constrained independent vector analysis source extraction algorithm for full-duplex communications. Simulation results confirm the effectiveness of the proposed method.

Index Terms— Acoustic echo cancellation, interference suppression, noise reduction, semi-blind source separation

1. INTRODUCTION

Acoustic echo cancellation (AEC), interference suppression, and noise reduction are fundamental tasks in full-duplex communications [1–3]. Consequently, numerous methods have been proposed over the past few decades to address these tasks either individually or jointly [4–10]. Among these, the adaptive filter followed by a post-beamformer has emerged as the most widely used approach [9]. However, in doubletalk scenarios where both far-end and near-end signals are active, the performance of such methods often degrades dramatically.

One way to enhance performance is by employing semi-blind source separation (SBSS) AEC algorithms [11–15]. While these algorithms are effective in handling doubletalk scenarios compared to traditional adaptive AEC algorithms, they are generally ineffective in dealing with near-end interference and noise. To address this issue, an offline joint AEC and source extraction algorithm based on independent vector analysis (IVA) was developed [16]. This algorithm shows promising performance when the near-end interference is relatively stationary compared to the target signal. However, in cases where the interference is non-stationary, there is a risk of extracting the interference instead of the target signal, known as permutation ambiguity [17].

In certain applications, the location of the target signal can be determined by a camera or estimated based on the observation signals. As shown in [18–22], such spatial prior information can be used to greatly enhance separation performance and address the permutation problem. With this in mind, this paper introduces a multichannel SBSS method, which consists of two stages. The first stage employs EISS to eliminate echo effects. Subsequently, we extend our previous geometrically constrained IVA (GC-IVA) approach [21] to handle an overdetermined situation for jointly suppressing interference and reducing noise. We henceforth refer to the proposed method as EISS-GCIVA. Simulations are conducted to validate the effectiveness of the proposed EISS-GCIVA system in doubletalk situations with non-stationary near-end interference and noise.

2. SIGNAL MODEL AND PROBLEM FORMULATION

We consider a full duplex communication scenario where an array of M microphones is used. The loudspeaker is relatively far from the array while the near-end target speaker and interference are close to the array. The m-th microphone signal can be denoted as

$$\begin{aligned} y_m(t) =& h_m(t) \star x(t) + a_{\mathrm{s},m}(t) \star s(t) \\ &+ \sum_{n=1}^{N} a_{\mathrm{e},m,n}(t) \star e_n(t) + v(t), \end{aligned} \tag{1}$$

where $\star$ denotes convolution operation, t is the time index, $h_m(t)$ is the room impulse response (RIR) from the loudspeaker to the m-th microphone, $x(t)$ is the far-end signal, $a_{\mathrm{s},m}(t)$ and $a_{\mathrm{e},m,n}(t)$ represent, respectively, the RIRs from target source and the n-th interference positions to the m-th microphone, $s(t)$, $e_n(t)$, $v(t)$ are the target signal, the n-th interference signal, and the stationary background noise, respectively. In this study, We consider only determined or over-determined situations, i.e., $M \geq 1 + N$. To make the deduced algorithms computationally efficient, a convolutive transfer function (CTF) model [23] is adopted, which expresses the signals given in (1) in the short-time-

Fourier-transform (STFT) domain as

$$Y_{m,i,j} = \sum_{l=0}^{L-1} H_{m,i,j-l} X_{i,j} + A_{\mathrm{s},m,i,j} S_{i,j} + \sum_{n=1}^{N} A_{\mathrm{e},m,n,i,j} E_{n,i,j} + V_{i,j}, \quad (2)$$

where i, j are, respectively, the frequency-bin and time-frame indices, $Y_{m,i,j}$, $X_{i,j}$, $S_{i,j}$, $E_{n,i,j}$ and $V_{i,j}$ are, respectively, the STFTs of $y_m(t)$, $x(t)$, $s(t)$, $e_n(t)$ and $v(t)$, and $H_{i,j-l}$, $A_{\mathrm{s},m,i,j}$, $A_{\mathrm{e},m,n,i,j}$ represent, respectively, $h_m(t)$, $a_{\mathrm{s},m}(t)$ and $a_{\mathrm{e},m,n}(t)$ in the STFT domain. The noisy near-end signal is defined as

$$Z_{m,i,j} = A_{\mathrm{s},m,i,j} S_{i,j} + \sum_{n=1}^{N} A_{\mathrm{e},m,n,i,j} E_{n,i,j} + V_{i,j}. \quad (3)$$

Without loss of generality, let us choose the first microphone as reference, then in matrix/vector form, (3) can be rewritten as

$$\mathbf{z}_{i,j} = \mathbf{a}_{\mathrm{s},i,j} A_{\mathrm{s},1,i,j} S_{i,j} + \sum_{n=1}^{N} \mathbf{a}_{\mathrm{e},n,i,j} A_{\mathrm{e},1,n,i,j} E_{n,i,j} + \mathbf{v}_{i,j}, \quad (4)$$

where

$$\begin{aligned} \mathbf{z}_{i,j} &= [Z_{1,i,j} \quad Z_{2,i,j} \quad \dots \quad Z_{M,i,j}]^T, \\ \mathbf{v}_{i,j} &= [V_{1,i,j} \quad V_{2,i,j} \quad \dots \quad V_{M,i,j}]^T, \\ \mathbf{a}_{\mathrm{s},i,j} &= \left[1 \quad \frac{A_{\mathrm{s},2,i,j}}{A_{\mathrm{s},1,i,j}} \quad \dots \quad \frac{A_{\mathrm{s},M,i,j}}{A_{\mathrm{s},1,i,j}}\right]^T, \\ \mathbf{a}_{\mathrm{e},n,i,j} &= \left[1 \quad \frac{A_{\mathrm{e},2,n,i,j}}{A_{\mathrm{e},1,n,i,j}} \quad \dots \quad \frac{A_{\mathrm{e},M,n,i,j}}{A_{\mathrm{e},1,n,i,j}}\right]^T. \end{aligned} \quad (5)$$

With the above CTF model, the problem in this study is to recover $A_{\mathrm{s},1,i,j} S_{i,j}$ given the multichannel observations and the reference signal $\mathbf{X}_{i,j}$.

3. CASCADED SEMI-BLIND SOURCE SEPARATION

In this section, we introduce a cascaded semi-blind-source-separation method comprising two stages: the first stage aims to eliminate the echo signal $\mathbf{z}_{i,j}$, while the second stage focuses on further suppressing near-end interference and noise.

3.1. AEC

In the first stage, the echo components can be expressed as

$$\tilde{\mathbf{y}}_{m,i,j} = \tilde{\mathbf{H}}_{m,i,j} \tilde{\mathbf{z}}_{m,i,j}, \quad (6)$$

where

$$\tilde{\mathbf{y}}_{m,i,j} = [Y_{m,i,j} \quad X_{i,j} \quad \cdots \quad X_{i,j-L+1}]^T, \quad (7)$$

$$\tilde{\mathbf{z}}_{m,i,j} = [Z_{m,i,j} \quad X_{i,j} \quad \cdots \quad X_{i,j-L+1}]^T, \quad (8)$$

$$\tilde{\mathbf{H}}_{m,i,j} = \begin{bmatrix} 1 & \mathbf{h}_{m,i,j}^T \\ \mathbf{0}_{L\times 1} & \mathbf{I}_L \end{bmatrix}, \quad (9)$$

with

$$\mathbf{h}_{m,i,j} = [H_{m,i,j} \quad H_{m,i,j-1} \quad \dots \quad H_{M,i,j-L+1}]^T. \quad (10)$$

The inverse process to estimate $\tilde{\mathbf{z}}_{m,i,j}$ can be written as

$$\tilde{\mathbf{z}}_{m,i,j} = \mathbf{W}_{m,i,j}^{\mathrm{AEC}} \tilde{\mathbf{y}}_{m,i,j}, \quad (11)$$

where

$$\mathbf{W}_{m,i,j}^{\mathrm{AEC}} = \begin{bmatrix} 1 & -\mathbf{h}_{m,i,j}^T \\ \mathbf{0}_L & \mathbf{I}_L. \end{bmatrix}. \quad (12)$$

let us define the near-end extraction filter

$$\mathbf{w}_{m,i,j}^{\mathrm{AEC}} = \begin{bmatrix} 1 & -\mathbf{h}_{m,i,j}^T \end{bmatrix}^H. \quad (13)$$

The near-end signal at the m-th microphone can then be estimated as

$$Z_{m,i,j} = \left(\mathbf{w}_{m,i,j}^{\mathrm{AEC}}\right)^H \tilde{\mathbf{y}}_{m,i,j}. \quad (14)$$

We model the near-end signal with a generalized Gaussian distribution (GCD), which is widely used in the literature of acoustic signal processing [24], i.e.,

$$p\left(\mathbf{z}_{m,j}\right) = \mathcal{N}_{\mathrm{GCD}}\left(\mathbf{z}_{m,j,}, \gamma_{\mathrm{z}}, \beta_{\mathrm{z}}\right) \propto \exp\left[-\left(\frac{\|\mathbf{z}_{m,j}\|_2}{\gamma_{\mathrm{z}}}\right)^{\beta_{\mathrm{z}}}\right],$$

where $||\cdot||_2$ denotes ℓ_2 norm, γ_{z} and β_{z} are two shape parameters, and

$$\mathbf{z}_{m,j} = [Z_{m,1,j} \quad Z_{m,2,j} \quad \dots \quad Z_{m,I,j}]^T \quad (15)$$

is a vector consisting of all frequency components of the near-end signal. To utilize the well known majorization-minimization (MM) method [25], we assume that $\gamma_{\mathrm{z}} > 0$, $0 < \beta_{\mathrm{z}} \leq 2$. Then, exploiting the mutual independence between the far- and near-end signals, one can write the following auxiliary function

$$\mathcal{L}_{m,j}^{\mathrm{AEC},+} = \sum_{i=1}^{I} \left(\mathbf{w}_{m,i,j}^{\mathrm{AEC}}\right)^H \mathbf{Q}_{m,i,j}^{\mathrm{AEC}} \mathbf{w}_{m,i,j}^{\mathrm{AEC}}, \quad (16)$$

where

$$\mathbf{Q}_{m,i,j}^{\mathrm{AEC}} = \alpha^{\mathrm{AEC}} \mathbf{Q}_{m,i,j-1}^{\mathrm{AEC}} + (1-\alpha^{\mathrm{AEC}}) \varphi(r_{\mathrm{z},m,j}) \tilde{\mathbf{y}}_{m,i,j} \tilde{\mathbf{y}}_{m,i,j}^H, \quad (17)$$

and

$$\varphi(r_{\mathrm{z},m,j}) = \left(r_{\mathrm{z},m,j}\right)^{\beta_{\mathrm{z}}-2}, \quad (18)$$

$$r_{\mathrm{z},m,j} = \sqrt{\textstyle\sum_{i=1}^{I} |\left(\mathbf{w}_{m,i,j}^{\mathrm{AEC}}\right)^H \tilde{\mathbf{y}}_{m,i,j}|^2}. \quad (19)$$

The element-wise source steering (EISS) is a computationally efficient algorithm to minimize (16). EISS update each element in $\mathbf{w}_{m,i,j}^{\mathrm{AEC}}$ individually with following update rule

$$W_{m,i,j,k}^{\mathrm{AEC}} \leftarrow W_{m,i,j-1,k}^{\mathrm{AEC}} - U_{m,i,j,k}^{\mathrm{AEC}}, \ k = 2, \dots, L+1, \quad (20)$$

where $W_{m,i,j,k}^{\mathrm{AEC}}$ is the m-th element of $\mathbf{w}_{m,i,j}^{\mathrm{AEC}}$ and $U_{m,i,j,k}^{\mathrm{AEC}}$ is known as the steering step size. Substituting (20) into (16), calculating the derivative with respect to $\left(U_{m,i,j,k}^{\mathrm{AEC}}\right)^*$ ($*$ denotes conjugate) and forcing the result to 0, we obtain

$$U_{m,i,j,k}^{\mathrm{AEC}} = 1 - \left[\left(\mathbf{w}_{m,i,j-1}^{\mathrm{AEC}}\right)^H \mathbf{Q}_{m,i,j}^{\mathrm{AEC}} \mathbf{w}_{m,i,j-1}^{\mathrm{AEC}}\right]^{-\frac{1}{2}}. \quad (21)$$

Applying EISS to each channel, we obtain the estimated $\mathbf{z}_{i,j}$.

3.2. Interference suppression and noise reduction

Now, we extend our previous work [21] to further suppress the near-near noise and interference after AEC. Similar with the overdetermined IVA (OverIVA) [25, 26], we assume that the rank of $E\left[\mathbf{v}_{i,j}\mathbf{v}_{i,j}^{H}\right]$ matrix is $M-1-N$. Then following demixing matrix is used to extract target signal

$$\mathbf{W}_{i,j}^{\mathrm{IVA}} = \begin{bmatrix} \widetilde{\mathbf{W}}_{i,j}^{\mathrm{IVA}} \\ \overline{\mathbf{W}}_{i,j}^{\mathrm{IVA}} \end{bmatrix}. \tag{22}$$

Since our goal is to separate the noise from the target and interference signals, we employ the specific structure of $\overline{\mathbf{W}}_{i,j}^{\mathrm{IVA}}$, corresponding to the OverIVA in this study. Specifically, we use:

$$\overline{\mathbf{W}}_{i,j}^{\mathrm{IVA}} = [\mathbf{\Phi}_{i,j} \ -\mathbf{I}_{M-1-N}], \tag{23}$$

where $\mathbf{\Phi}_{i,j}$ contains adjustable parameters. Without loss of generality, we choose the first row in $\widetilde{\mathbf{W}}_{i,j}^{\mathrm{IVA}}$, i.e., $\widetilde{\mathbf{w}}_{1,i,j}^{\mathrm{IVA}}$, as the target extraction filter. To incorporate spatial prior information about the target and interference signal, we assume that

$$p\left(\widetilde{\mathbf{w}}_{n',i,j}^{\mathrm{IVA}}\right) \propto \exp\left(-\sum_{i=1}^{I}\lambda_{n',j}\left\|\widetilde{\mathbf{w}}_{n',i,j}^{\mathrm{IVA}} - \mathbf{d}_{\hat{\theta},i}\right\|_2^2\right), \tag{24}$$

where $\lambda_{n',j}$ is the weighting parameter,

$$\mathbf{d}_{\hat{\theta},i} = \begin{bmatrix} 1 & e^{-j\omega_i\tau_{2,\hat{\theta}}} & \dots & e^{-j\omega_i\tau_{M,\hat{\theta}}} \end{bmatrix}^T \tag{25}$$

is a steering vector and ω_i is the angular frequency corresponding to i-th frequency bin and $n' = 1, 2, \dots, N+1$. The time difference of arrival of the signal incident from $\hat{\theta}$ between the m-th and first microphones is denoted as $\tau_{m,\hat{\theta}}$. To use the MM method, we model the target signal and interference with a GCD as well, i.e., $p(\mathbf{s}_j) = \mathcal{N}_{\mathrm{GCD}}(\mathbf{s}_j, \gamma_1, \beta_1)$ and $p(\mathbf{e}_{n,j}) = \mathcal{N}_{\mathrm{GCD}}(\mathbf{e}_{n,j}, \gamma_{n+1}, \beta_{n+1})$ where

$$\mathbf{s}_j = [S_{1,j} \quad S_{2,j} \quad \dots \quad S_{I,j}]^T, \tag{26}$$

$$\mathbf{e}_{n,j} = [E_{n,1,j} \quad E_{n,2,j} \quad \dots \quad E_{n,I,j}]^T. \tag{27}$$

By adopting the maximum *a posteriori* (MAP) and MM method [27], we can derive the following recursive cost function

$$\mathcal{L}_j^{\mathrm{IVA}}\left(\widetilde{\mathbf{W}}_{i,j}^{\mathrm{IVA}}\right) = \sum_{n'=1}^{N+1}\Big[\left(\tilde{\mathbf{w}}_{n',i,j}^{\mathrm{IVA}}\right)^H \mathbf{Q}_{n',i,j}^{\mathrm{IVA}}\tilde{\mathbf{w}}_{n',i,j}^{\mathrm{IVA}} + \lambda_{n',j}||\tilde{\mathbf{w}}_{n',i,j}^{\mathrm{IVA}} - \mathbf{d}_{\hat{\theta}_{n'},i}||_2^2\Big] - 2\sum_{i=1}^{I}\log\left|\det\mathbf{W}_{i,j}^{\mathrm{IVA}}\right|, \tag{28, 29}$$

where $\left(\tilde{\mathbf{w}}_{n',i,j}^{\mathrm{IVA}}\right)^H$ is the n'-th row of $\widetilde{\mathbf{W}}_{i,j}^{\mathrm{IVA}}$ and

$$\mathbf{Q}_{n',i,j}^{\mathrm{IVA}} = \alpha^{\mathrm{IVA}}\mathbf{Q}_{n',i,j-1}^{\mathrm{IVA}} + (1-\alpha^{\mathrm{IVA}})\varphi(r_{n',j})\mathbf{z}_{i,j}\mathbf{z}_{i,j}^H, \tag{30}$$

with

$$\varphi(r_{n',j}) = (r_{n',j})^{\beta_{n'}-2}, \tag{31}$$

$$r_{n',j} = \sqrt{\sum_{i=1}^{I}\left|\left(\mathbf{w}_{n',i,j}^{\mathrm{IVA}}\right)^H\mathbf{z}_{i,j}\right|^2}. \tag{32}$$

3.2.1. Update of $\widetilde{\mathbf{W}}_{i,j}^{\mathrm{IVA}}$

Now, we estimate the demixing matrix with the geometrically constrained iterative source steering (GC-ISS) update rules, which are

$$\widetilde{\mathbf{W}}_{i,j}^{\mathrm{IVA}} = \widetilde{\mathbf{W}}_{i,j-1}^{\mathrm{IVA}} - \mathbf{u}_{m,i,j}^{\mathrm{IVA}}\left(\mathbf{w}_{m,i,j-1}^{\mathrm{IVA}}\right)^H, \tag{33}$$

where

$$\mathbf{u}_{m,i,j}^{\mathrm{IVA}} = \begin{bmatrix} U_{m,i,j,1}^{\mathrm{IVA}} & U_{m,i,j,2}^{\mathrm{IVA}} & \dots & U_{m,i,j,N+1}^{\mathrm{IVA}} \end{bmatrix}^T. \tag{34}$$

For each pair of m and n', if $n' \neq m$, the optimal estimate of $U_{m,i,j,n'}^{\mathrm{IVA}}$ is

$$\hat{U}_{m,i,j,n'}^{\mathrm{IVA}} = \frac{G_{m,i,j,n'} + \lambda_{n',j}\tilde{\mathbf{d}}_{\hat{\theta}_{n',i,j-1}}^H\mathbf{w}_{m,i,j-1}^{\mathrm{IVA}}}{\tilde{G}_{m,i,j,n'} + \lambda_{n',j}\left(\mathbf{w}_{m,i,j-1}^{\mathrm{IVA}}\right)^H\mathbf{w}_{m,i,j-1}^{\mathrm{IVA}}}, \tag{35}$$

where

$$G_{m,i,j,n'} = \left(\mathbf{w}_{n',i,j-1}^{\mathrm{IVA}}\right)^H\mathbf{Q}_{n',i,j}^{\mathrm{IVA}}\mathbf{w}_{m,i,j-1}^{\mathrm{IVA}}, \tag{36}$$

$$\tilde{G}_{m,i,j,n'} = \left(\mathbf{w}_{m,i,j-1}^{\mathrm{IVA}}\right)^H\mathbf{Q}_{n',i,j}^{\mathrm{IVA}}\mathbf{w}_{m,i,j-1}^{\mathrm{IVA}}, \tag{37}$$

$$\tilde{\mathbf{d}}_{\hat{\theta}_{n',i,j-1}} = \left(\mathbf{w}_{n',i,j-1}^{\mathrm{IVA}}\right) - \mathbf{d}_{\hat{\theta}_{n',i}}, \tag{38}$$

while if $n' = m$, $U_{m,i,j,n'}^{\mathrm{IVA}}$ is given as

$$\hat{U}_{m,i,j,n'}^{\mathrm{IVA}} = \begin{cases} 1 - \eta_{i,j}\frac{|\eta_{i,j}| + \sqrt{|\eta_{i,j}|^2 + 4\zeta_{i,j}}}{2\zeta_{i,j}|\eta_{i,j}|}, & (\text{if } \eta_i \neq 0) \\ 1 - \frac{1}{\sqrt{\zeta_{i,j}}}, & (\text{else}) \end{cases} \tag{39}$$

where

$$\eta_{i,j} = \lambda_{m,j}\mathbf{d}_{\hat{\theta}_m,i}^H\mathbf{w}_{m,i,j-1}^{\mathrm{IVA}}, \tag{40}$$

$$\zeta_{i,j} = \tilde{G}_{m,i,j,m} + \lambda_{m,j}\left(\mathbf{w}_{m,i,j-1}^{\mathrm{IVA}}\right)^H\mathbf{w}_{m,i,j-1}^{\mathrm{IVA}}. \tag{41}$$

3.2.2. Update of $\mathbf{\Phi}_{i,j}$

After updating the first $N+1$ rows, the noise associated matrix is updated as

$$\mathbf{\Phi}_{i,j} = \left[\mathbf{E}_2\mathbf{C}_{i,j}^{\mathrm{IVA}}\left(\widetilde{\mathbf{W}}_{i,j}^{\mathrm{IVA}}\right)^H\right]\left[\mathbf{E}_1\mathbf{C}_{i,j}^{\mathrm{IVA}}\left(\widetilde{\mathbf{W}}_{i,j}^{\mathrm{IVA}}\right)^H\right]^{-1}, \tag{42}$$

where

$$\mathbf{C}_{i,j}^{\mathrm{IVA}} = \alpha^{\mathrm{IVA}}\mathbf{C}_{i,j-1}^{\mathrm{IVA}} + (1-\alpha^{\mathrm{IVA}})\mathbf{z}_{i,j}\mathbf{z}_{i,j}^H, \tag{43}$$

$$\mathbf{E}_1 = \begin{bmatrix} \mathbf{I}_N & \mathbf{0}_{N\times(M-N)} \end{bmatrix}, \tag{44}$$

$$\mathbf{E}_2 = \begin{bmatrix} \mathbf{0}_{(M-N)\times N} & \mathbf{I}_{M-N} \end{bmatrix}. \tag{45}$$

After updating the whole system, the near-end target signal will be extracted as

$$\hat{S}_{i,j} = \left(\tilde{\mathbf{w}}_{1,i,j}^{\mathrm{IVA}}\right)^H\mathbf{z}_{i,j}. \tag{46}$$

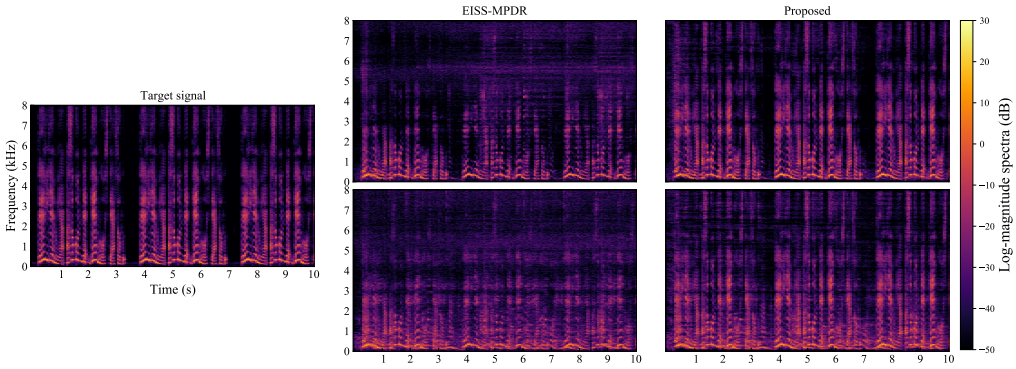

Fig. 1. Spectrograms of the extracted target signals. The first row corresponds to the target signals extracted in the first case. The second row corresponds to the extracted target signals in the second case.

4. SIMULATION

In this section, we evaluate the performance of the proposed EISS-GCIVA system. It is important to note that the multichannel state space method introduced in [28] cannot handle near-end interference and noise suppression. Additionally, the joint blind source extraction and AEC algorithm discussed in [16] is offline. Therefore, to demonstrate the near-end noise reduction, interference suppression, and near-end signal preservation capabilities of the proposed method, we employ EISS with a post minimal power distortionless response (MPDR) beamformer as the baseline.

A room with dimensions of $8\,\mathrm{m}\times 8\,\mathrm{m}\times 3\,\mathrm{m}$ is simulated based on the image-source method [29]. For convenience, a Cartesian coordinate system is adopted, with the left-bottom corner of the room serving as the origin. A uniform linear microphone array consisting of 8 microphones, spaced $4\,\mathrm{cm}$ apart, is employed. The array is oriented parallel to the x-axis, with its center positioned at $(4\,\mathrm{m},\, 4\,\mathrm{m},\, 1\,\mathrm{m})$. A loudspeaker is positioned at $(1\,\mathrm{m},\, 4\,\mathrm{m},\, 2.5\,\mathrm{m})$. Additionally, a target speaker is located at $(4.77\,\mathrm{m},\, 4.64\,\mathrm{m},\, 1.51\,\mathrm{m})$, while an interference source is situated at $(4\,\mathrm{m},\, 5.2\,\mathrm{m},\, 1.66\,\mathrm{m})$. Three spatially distributed white Gaussian noise sources are randomly placed in the room, ensuring a minimum distance of $2\,\mathrm{m}$ from the microphone array.

We explore two scenarios. In the first case, we examine a less challenging setting where there is no interference present and the reverberation time t_{60} is set to $200\,\mathrm{ms}$. For the second scenario, we consider a more challenging environment with t_{60} set to $400\,\mathrm{ms}$, and the interference consists of a 10-second signal taken from the CHiME-3 Caf noise [30]. In both cases, the signal-to-echo ratio (SER) is maintained at $0\,\mathrm{dB}$, the signal-to-noise ratio (SNR) at $20\,\mathrm{dB}$, and both far- and near-end signals consist of 10-second-long speech signals from two female speakers. For the second scenario, the signal-to-interference ratio (SIR) is set to $10\,\mathrm{dB}$. A 512-sample Von Hann window with $75\,\%$ overlap is utilized, resulting in an algorithmic delay of $32\,\mathrm{ms}$. The shape parameters β_{z} for the near-end mixture, β_1 for the target signal, and β_2 for interference are all set to 0.4. The forgetting factor α^{AEC} is set to 0.992, while α^{IVA} is set to 0.98. The initial values of $\lambda_{n',0}$ are all set to 1 and decay during iterations according to $\lambda_{n',j} = \max\left(0.001, 0.8^j \times \lambda_{n',0}\right)$. To mimic realistic conditions, biased direction estimates are provided. Specifically, the estimate of the target incidence angle is 20° when its actual angle is 40°, and the estimate of the interference incidence angle is 100° while it is actually 80°. We assess all the studied algorithms based on their output signal-to-interference-and-noise ratio (oSINR), output SER (oSER), and true return echo loss enhancement (tERLE) [11].

Table 1. Performance of the compared methods in the first case.

Systems	Case	oSINR	oSER	tERLE
EISS-MPDR	1	25.13	6.95	10.33
	2	20.79	4.69	8.26
EISS-GCIVA	1	33.19	9.74	10.30
	2	24.11	7.13	8.29

The results are presented in Table 1. It is seen that EISS-GCIVA achieves a higher oSINR compared to EISS-MPDR, indicating the superior near-end noise and interference suppression capabilities of the proposed method. Additionally, it is observed that EISS-MPDR demonstrates similar tERLE performance to EISS-GCIVA. This similarity arises because the echo is primarily estimated in the first stage. However, the oSER of EISS-GCIVA is significantly higher than that of EISS-MPDR. To illustrate this difference, we plot the spectrograms of the extracted target signals in Fig. 1. It is apparent that the target signal extracted by EISS-MPDR is distorted, whereas the quality of the target signal extracted by EISS-GCIVA is significantly superior, highlighting the effectiveness of the proposed method in preserving the target signal.

5. CONCLUSION

Effective joint acoustic echo cancellation and interference and noise suppression play a critical role in full-duplex communications. In this study, we introduced a cascaded semi-blind source separation method designed to eliminate echo while simultaneously suppressing near-end interference and noise in double-talk scenarios with minimal algorithmic delay. Simulation results confirmed the superior performance of the proposed method.

6. REFERENCES

[1] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. IEEE ICASSP*, 1988, pp. 2570–2573.

[2] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in network and acoustic echo cancellation*. Berlin, Germany: Springer-Verlag, 2001.

[3] Y. Huang, J. Chen, and J. Benesty, "Immersive audio schemes," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 20–32, Jan. 2011.

[4] E. Ferrara, "Fast implementations of LMS adaptive filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 474–475, Aug. 1980.

[5] G. Long, F. Ling, and J. G. Proakis, "The LMS algorithm with delayed coefficient adaptation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 9, pp. 1397–1405, Sept. 1989.

[6] J.-I. Nagumo and A. Noda, "A learning method for system identification," *IEEE Trans. Autom. Control*, vol. 12, no. 3, pp. 282–287, Jun. 1967.

[7] C. Paleologu, J. Benesty, and S. Ciochină, "An improved proportionate NLMS algorithm based on the l_0 norm," in *Proc. IEEE ICASSP*, 2010, pp. 309–312.

[8] V Panuska, "An adaptive recursive-least-squares identification algorithm," in *Proc. IEEE ICASSP*, 1969, pp. 65–65.

[9] W. Kellermann, "Acoustic echo cancellation for beamforming microphone arrays," in *M*icrophone Arrays, D. Ward M. Brandstein, Ed., pp. 281–306. Springer, 2001.

[10] C. Paleologu, J. Benesty, and S. Ciochină, "A robust variable forgetting factor recursive least-squares algorithm for system identification," *IEEE Signal Process. Lett.*, vol. 15, pp. 597–600, Oct. 2008.

[11] F. Nesta, T. S. Wada, and B.-H. Juang, "Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 583–599, Mar. 2011.

[12] W. Wang, Y. Na, Z. Liu, B. Tian, and Q. Fu, "Weighted recursive least square filter and neural network based residual echo suppression for the AEC-challenge," in *Proc. IEEE ICASSP*, 2021, pp. 141–145.

[13] G. Cheng, L. Liao, H. Chen, and J. Lu, "Semi-blind source separation for nonlinear acoustic echo cancellation," *IEEE Signal Process. Lett.*, vol. 28, pp. 474–478, Feb. 2021.

[14] G. Cheng, L. Liao, K. Chen, Y. Hu, C. Zhu, and J. Lu, "Semi-blind source separation using convolutive transfer function for nonlinear acoustic echo cancellation," *J. Acoust. Soc. Am.*, vol. 153, no. 1, pp. 88–95, 2023.

[15] K. Lu, X. Wang, T. Ueda, S. Makino, and J. Chen, "A computationally efficient semi-blind source separation approach for nonlinear echo cancellation based on an element-wise iterative source steering," in *Proc. IEEE ICASSP*, 2024, pp. 756–760.

[16] T. Haubner, W. Kellermann, and Z. Koldovský, "Joint acoustic echo cancellation and blind source extraction based on indepenednt vector extraction," in *Proc IWAENC*, 2022.

[17] S. Makino, *Audio source separation*. Switzerland: Springer, 2018.

[18] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech, Audio Process.*, vol. 10, no. 6, pp. 352–562, Dec. 2002.

[19] L. Li and K. Koishida, "Geometrically constrained independent vector analysis for directional speech enhancement," in *Proc. IEEE ICASSP*, 2020, pp. 846–850.

[20] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 3545–3558, Jun. 2020.

[21] K. Mo, X. Wang, Y. Yang, T. Ueda, S. Makino, J. Chen, "On joint dereverberation and source separation with geometrical constraints and iterative source steering," in *Proc APSIPA ASC*, 2023.

[22] X. Wang, A. Brendel, G. Huang, Y. Yang, W. Kellermann, and J. Chen, "Spatially informed independent vector analysis for source extraction based on the convolutive transfer function model," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[23] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, 2009.

[24] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[25] R. Scheibler and N. Ono, "MM algorithms for joint independent subspace analysis with application to blind single and multi-source extraction," *arxiv*, 2020.

[26] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc WASPAA*, pp. 185–189, 10 2019.

[27] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans on Signal Process.*, vol. 68, pp. 3545–3558, 2020.

[28] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2065–2079, 2012.

[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[30] J. Barker, R. Marxer, E. Vincent, and A. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 5210–5214.