Heavy-tailed Distributions-Based Online Semi-blind Source Separation for Nonlinear Echo Cancellation

Liyuan Zhang*, Xianrui Wang*†, Yichen Yang*†, Tetsuya Ueda*, Shoji Makino* and Jingdong Chen†
* Waseda University, Japan
† Northwestern Polytechnical University, Xi'an, China

***Abstract*—Recently proposed semi-blind source separation (SBSS) based acoustic echo cancellation (AEC) algorithms have attracted significant research interest due to their ability to track dynamic acoustic environments in the presence of near-end signals. Two source models are considered in these existing algorithms, i.e., the spherical generalized super-Gaussian distribution and the circular super-Gaussian distribution with a low-rank spectrogram model. In this paper, we aim to further enhance AEC performance by leveraging more flexible source models. Several novel algorithms are subsequently proposed. Simulations demonstrate the superiority of proposed algorithms in various situations.**

I. Introduction

Acoustic echo cancellation (AEC) plays an indispensable role in real-time full-duplex communication [1]–[3]. The performance of conventional adaptive filters based algorithms generally suffers dramatic degradation in double-talk situations, i.e., when far-end signal and near-end signal coexist [1], [4], [5]. To address this problem, researchers reformulated AEC as a semi-blind source separation (SBSS) problem [6], [7] and subsequently derived several SBSS-AEC algorithms [8]–[12].

Existing SBSS-AEC algorithms can be roughly divided into two categories based on the source model assumption: the auxiliary function-based independent vector analysis (AuxIVA) [13] which considers spherical super-Gaussian distribution [11], [12], and the Itakura-Saito divergence-based independent low-rank matrix analysis (IS-ILRMA) [14], [15], which considers circular super-Gaussian distribution and non-negative matrix factorization (NMF) spectrogram model [10].

Although these algorithms have achieved satisfying AEC performance, there is still room for further improvement by considering more generalized and flexible source models. In this paper, we adopt two heavy-tailed distributions: the complex generalized super-Gaussian distribution (GGD) and Student's t-distribution. Subsequently, we derive three novel algorithms, namely, GGD-ILRMA, t-AuxIVA and t-ILRMA based SBSS-AEC. To update NMF parameters in a real-time manner, we apply recursive approximation and semi-supervised NMF (SSNMF) techniques [16], [17]. The performance of the proposed algorithms is studied under complex real-time circumstances. Several simulations validate the efficacy and superiority of the proposed algorithms.

The remainder of this paper is organized as follows: In Section II, a bilinear signal model and the problem formulation of SBSS-AEC are introduced. Next, on the basis of conventional GGD-AuxIVA and IS-ILRMA based SBSS-AEC [10]–[12] described in Section III-A, we propose GGD-ILRMA-based SBSS-AEC algorithm utilizing alternating iterative projection (AIP) as update rules and recursive approximation for online implementation in Section III-B. In Section III-C, all the mentioned GGD-based SBSS-AEC algorithms are extended to their heavy-tailed counterparts with Student's t-distributions, where t-AuxIVA and t-ILRMA based SBSS-AEC algorithms are proposed. The feasibility of SSNMF is discussed in Section III-D. Simulations and results considering two scenarios, i.e., speech and music, are shown and analyzed in Section IV.

II. Signal Model and Problem Formulation

Considering a full-duplex communication scenario, a loudspeaker with unknown non-linearity is used to play the far-end signal. The output of the loudspeaker is convolved with acoustic impulse response (AIR) to generate the echo. The superimposition of the near-end signal and the echo is captured by the microphone and then transmitted back to the far end. We use odd power series expansion technique to approximate the non-linearity of the loudspeaker as described in [9], [18], [19]. In double-talk situation, the observed microphone signal in short-time Fourier transform (STFT) domain can be denoted as [10]–[12]

$$\begin{aligned} Y_{i,j} &= E_{i,j} + S_{i,j} \\ &= \sum_{n=1}^{N} \sum_{l=1}^{L} a_n H_{i,j,l} X_{n,i,j-l+1} + S_{i,j}, \end{aligned} \tag{1}$$

where $i = 1, \ldots, I$ and $j = 1, \ldots, J$ are frequency-bin and time-frame indices, respectively, I and J are the number of frequency bins and time frames, respectively, N is the expansion order, a_n is the n-th order odd power series expansion coefficient, L is the convolutive transfer function (CTF) filter length, $H_{i,j,l}$ is the CTF filter coefficient [10], [20], and $X_{n,i,j}$, $E_{i,j}$ and $S_{i,j}$ are the STFTs of nth-order far-end signal, nonlinear echo and near-end signal, respectively.

The nonlinear echo can be reformulated as a bilinear form [11]

$$E_{i,j} = \mathbf{h}_{i,j}^T \mathbf{X}_{i,j} \mathbf{a}, \tag{2}$$

where

$$\mathbf{h}_{i,j} = [H_{i,j,1} \quad H_{i,j,2} \quad \cdots \quad H_{i,j,L}]^T, \tag{3}$$

$$\mathbf{a} = [a_1 \quad a_2 \quad \cdots \quad a_N]^T, \tag{4}$$

$$\mathbf{X}_{i,j} = [\tilde{\mathbf{x}}_{i,j} \quad \tilde{\mathbf{x}}_{i,j-1} \quad \cdots \quad \tilde{\mathbf{x}}_{i,j-L+1}]^T, \tag{5}$$

$$\tilde{\mathbf{x}}_{i,j} = [X_{1,i,j} \quad X_{2,i,j} \quad \cdots \quad X_{N,i,j}]^T, \tag{6}$$

$\mathbf{X}_{i,j} \in \mathbb{C}^{L\times N}$ is a matrix constructed by far-end signal series $X_{n,i,j}$ and $(\cdot)^T$ denotes the transpose operation. Using (2), the estimation of a_n and $H_{i,j,l}$ can be interpreted as two sub-separate SBSS problems. After defining a new far-end reference vector

$$\mathbf{x}_{i,j} = \mathbf{X}_{i,j}\mathbf{a}, \tag{7}$$

(2) is rewritten as

$$E_{i,j} = \mathbf{h}_{i,j}^T \mathbf{x}_{i,j}. \tag{8}$$

Then, the mixing process in (1) can be rewritten with a vector form

$$\tilde{\mathbf{y}}_{i,j} = \mathbf{H}_{i,j}\tilde{\mathbf{s}}_{i,j}, \tag{9}$$

with

$$\tilde{\mathbf{y}}_{i,j} = \begin{bmatrix} Y_{i,j} & \mathbf{x}_{i,j}^T \end{bmatrix}^T, \tag{10}$$

$$\tilde{\mathbf{s}}_{i,j} = \begin{bmatrix} S_{i,j} & \mathbf{x}_{i,j}^T \end{bmatrix}^T, \tag{11}$$

$$\mathbf{H}_{i,j} = \begin{bmatrix} 1 & \mathbf{h}_{i,j}^T \\ \mathbf{0}_{L\times 1} & \mathbf{I}_L \end{bmatrix}, \tag{12}$$

where $\mathbf{H}_{i,j} \in \mathbb{C}^{(L+1)\times(L+1)}$ is the mixing matrix, $\mathbf{0}_{L\times 1}$ is a column vector of length L with all elements equal to 0, $\mathbf{I}_L$ is an identity matrix of size $L \times L$. Assuming that $\mathbf{H}_{i,j}$ is non-singular, to extract the near-end signal, the demixing matrix $\hat{\mathbf{W}}_{i,j} \in \mathbb{C}^{(L+1)\times(L+1)}$ is obtained as the inverse of $\mathbf{H}_{i,j}$, which is

$$\hat{\mathbf{W}}_{i,j} = \begin{bmatrix} 1 & -\hat{\mathbf{h}}_{i,j}^T \\ \mathbf{0}_{L\times 1} & \mathbf{I}_L \end{bmatrix}, \tag{13}$$

where $\hat{\mathbf{h}}_{i,j}^T$ is a column vector with L parameters to be estimated. With this structure, the near-end signal is extracted as

$$\hat{S}_{i,j} = \hat{\mathbf{w}}_{i,j}^H \tilde{\mathbf{y}}_{i,j}, \tag{14}$$

where the extraction filter $\hat{\mathbf{w}}_{i,j} = [1 \quad -\hat{\mathbf{h}}_{i,j}^T]^H$ is the first row of (13) and $(\cdot)^H$ denotes the Hermitan transpose. Therefore, the target of SBSS-AEC in double-talk situation is transformed into estimating $\hat{\mathbf{w}}_{i,j}$ to extract the near-end signal. Note that (2) can also be written as $E_{i,j} = \mathbf{a}^T \mathbf{x}_{i,j}^\star$, where $\mathbf{x}_{i,j}^\star = \mathbf{X}_{i,j}^T \mathbf{h}_{i,j}$. Similar derivation as (9)–(14) is omitted for simplification [11].

III. Generalization of Online SBSS-AEC Algorithms

A. Conventional method

In SBSS-AEC, the near-end source can be extracted by exploiting mutual independence between near-end and reference signals. Therefore, since the determinant of (13) equals one, the following recursive negative log-likelihood function is derived [12], [16]

$$\mathcal{L}_j = -\frac{1}{\sum_{j'=1}^{j}(\eta)^{j-j'}} \sum_{j'=1}^{j} (\eta)^{j-j'} \log p\left(\mathbf{s}_{j'}\right), \tag{15}$$

where $\eta \in (0,1)$ is a forgetting factor and

$$\mathbf{s}_j = [S_{1,j} \quad S_{2,j} \quad \ldots \quad S_{I,j}]^T. \tag{16}$$

In GGD-AuxIVA-based SBSS-AEC, the near-end signal is modeled with a spherical complex GGD [10]–[12], [21]

$$p_{\mathrm{GGD}}(\mathbf{s}_j) \propto \exp\left[-\left(\frac{\|\mathbf{s}_j\|_2}{\gamma}\right)^{\beta}\right], \tag{17}$$

where $\|\cdot\|_2$ stands for ℓ_2 norm, γ and β stand for positive scale and shape parameters, respectively. We assume that $0 < \beta \leq 2$ to satisfy the precondition of majorization-minimization (MM) method [21], [22].

Here, a_n and $H_{i,j,l}$ are updated alternatingly. When we fix a_n, the following cost function for CTF coefficients $\hat{\mathbf{h}}_{i,j}$ can be obtained using MM method

$$\mathcal{L}_j^{\mathbf{h},+} = \sum_{i=1}^{I} \hat{\mathbf{w}}_{i,j}^H \mathbf{V}_{i,j} \hat{\mathbf{w}}_{i,j}. \tag{18}$$

To further reduce computational cost, we employ AIP update rules [11], where the auxiliary matrix $\mathbf{V}_{i,j} \in \mathbb{C}^{(L+1)\times(L+1)}$ is partitioned as

$$\mathbf{V}_{i,j} = \begin{bmatrix} \sigma_{y,i,j}^2 & \mathbf{g}_{i,j}^H \\ \mathbf{g}_{i,j} & \mathbf{C}_{i,j} \end{bmatrix}, \tag{19}$$

$\mathbf{g}_{i,j}$ and $\mathbf{C}_{i,j}$ are recursively updated based on (18)

$$\sigma_{y,i,j}^2 = \eta\sigma_{y,i,j-1}^2 + (1-\eta)\varphi_{\mathrm{GGD}}(\sigma_{s,j})\left|Y_{i,j}\right|^2, \tag{20}$$

$$\mathbf{g}_{i,j} = \eta\mathbf{g}_{i,j-1} + (1-\eta)\varphi_{\mathrm{GGD}}(\sigma_{s,j}) Y_{i,j}^* \mathbf{x}_{i,j}, \tag{21}$$

$$\mathbf{C}_{i,j} = \eta\mathbf{C}_{i,j-1} + (1-\eta)\varphi_{\mathrm{GGD}}(\sigma_{s,j}) \mathbf{x}_{i,j}\mathbf{x}_{i,j}^H, \tag{22}$$

where $(\cdot)^*$ denotes conjucate operation, $\varphi_{\mathrm{GGD}}(\sigma_{s,j})$ is the score function

$$\varphi_{\mathrm{GGD}}(\sigma_{s,j}) = \sigma_{s,j}^{\beta-2}, \tag{23}$$

and $\sigma_{s,j} = \|\hat{\mathbf{s}}_j\|_2$ is the auxiliary variable. Next, the CTF filter is updated as [11]

$$\hat{\mathbf{h}}_{i,j} = \left(\mathbf{C}_{i,j}^{-1}\mathbf{g}_{i,j}\right)^*. \tag{24}$$

When we fix $H_{i,j,l}$, the update of expansion coefficient $\hat{\mathbf{a}}$ can be derived similarly as (18)–(24) [11]. After updating all the parameters, the near-end signal can be extracted using (14).

While GGD-AuxIVA-based SBSS-AEC is effective in most scenarios, as shown in (17), the multivariate source model of AuxIVA is inflexible and cannot cope with specific harmonic structures of each source due to higher-order correlations between frequency bins [23]. As another state-of-the-art BSS algorithm, ILRMA is more generalized and effective in modeling sources by using the low-rank time-frequency NMF structure, especially in music scenarios [14]. Nevertheless, the ILRMA-based SBSS-AEC method in [10] assumes IS-NMF as the generative model, the probabilistic distribution of which is deficient compared with GGD-NMF. Therefore, in this paper, we first improve the performance of GGD-AuxIVA-based and IS-ILRMA-based SBSS-AEC by using GGD-NMF as the source generative model and propose a new GGD-

ILRMA based SBSS-AEC algorithm.

B. Proposed GGD-ILRMA based SBSS-AEC

In GGD-ILRMA-based SBSS-AEC, the source generative model follows a time-frequency-wise isotropic complex GGD

$$p_{\mathrm{GGD}}(S_{i,j}) = \frac{\beta}{2\pi r_{i,j}^2 \Gamma(\frac{2}{\beta})} \exp\left[-\left(\frac{|S_{i,j}|}{r_{i,j}} \right)^{\beta} \right], \tag{25}$$

$$r_{i,j}^{p} = \sum_{k=1}^{K} t_{i,k} v_{k,j}, \tag{26}$$

where $t_{i,k} \geq 0$ and $v_{k,j} \geq 0$ are the nonnegative basis and activation elements of the basis matrix $T_n \in \mathbb{R}_{\geq 0}^{I \times K}$ and the activation matrix $V_n \in \mathbb{R}_{\geq 0}^{K \times J}$, repectively, $k = 1, \ldots, K$ is the integral index of the basis, K is the number of NMF bases, $r_{i,j}$ is a time-frequency-varying variance corresponding to the low-rank source model, $p > 0$ is the parameter that defines NMF domain and $\Gamma(\cdot)$ is the Gamma function [24]. Using MM method, the score function is derived as

$$\varphi_{\mathrm{GGD}}(r_{i,j}) = \frac{\beta}{2} \left| \hat{S}_{i,j} \right|^{\beta - 2} r_{i,j}^{-\beta}. \tag{27}$$

In ILRMA, the update rules for $\hat{\mathbf{w}}_{i,j}$ are equivalent to those in AIP-based AuxIVA [11], [14]. Regarding the source model, we adopt the update rules in literature [24] and adapt the update of basis elements from offline to online scenarios using recursive approximation [10], [16]. Information from the previous frames is utilized to update the scaling factor of the basis elements in the current frame:

$$t_{i,k}^{\mathrm{GGD}} \leftarrow t_{i,k}^{\mathrm{GGD}} \left[\frac{\beta \sum_{j'=1}^{j} \alpha_{\mathrm{GGD}}^{j-j'} \left| \hat{S}_{i,j'} \right|^{\beta} r_{i,j'}^{-(p+\beta)} v_{k,j'}^{\mathrm{GGD}}}{2 \sum_{j'=1}^{j} \alpha_{\mathrm{GGD}}^{j-j'} r_{i,j'}^{-p} v_{k,j'}^{\mathrm{GGD}}} \right]^{\frac{p}{p+\beta}}, \tag{28}$$

$$v_{k,j}^{\mathrm{GGD}} \leftarrow v_{k,j}^{\mathrm{GGD}} \left[\frac{\beta \sum_{i=1}^{I} \left| \hat{S}_{i,j} \right|^{\beta} r_{i,j}^{-(p+\beta)} t_{i,k}^{\mathrm{GGD}}}{2 \sum_{i=1}^{I} r_{i,j}^{-p} t_{i,k}^{\mathrm{GGD}}} \right]^{\frac{p}{p+\beta}}, \tag{29}$$

where $\alpha_{\mathrm{GGD}} \in (0, 1)$ is a forgetting factor for NMF variables. With an appropriate setup of α_{GGD}, the convergence speed of the NMF model can be increased during the early stages.

C. Proposed Student's t-distribution-based SBSS-AEC

Student's t-distribution is also a popular heavy-tailed distribution model, which is effective in modeling audio sources due to its heavier tail and controllable degrees of freedom parameter [25]–[27]. In this paper, we extend GGD-AuxIVA-based and GGD-ILRMA-based SBSS-AEC using Student's t-distributions, and propose t-AuxIVA-based and t-ILRMA-based SBSS-AEC.

For t-AuxIVA-based SBSS-AEC, we assume a spherical complex Student's t-distribution with zero-mean as p.d.f. of the near-end signal [25], [26]:

$$p_{\mathrm{t}}(\mathbf{s}_j) \propto \left(1 + \frac{2}{v} \frac{\|\mathbf{s}_j\|_2^2}{\gamma_j} \right)^{-\frac{I+v}{2}}, \tag{30}$$

where $v > 0$ is the degrees of freedom parameter and γ_j is the uniform variance over frequency bins in the j-th time frame. Using (15) and (30), the following score function is derived:

$$\varphi_{\mathrm{t}}(\sigma_{s,j}) = \left(1 + \frac{I}{v} \right) \left(\frac{\gamma_j}{2} + \frac{\sigma_{s,j}^2}{v} \right)^{-1}. \tag{31}$$

Since the cost function is identical to (18) when we fix a_n, $\hat{\mathbf{w}}_{i,j}$ can be updated with the AIP algorithm (19)–(22).

As to t-ILRMA-based SBSS-AEC, the isotropic complex Student's t-distribution is assumed as the following source generative model:

$$p_{\mathrm{t}}(S_{i,j}) = \frac{1}{\pi r_{i,j}^2} \left(1 + \frac{2}{v} \frac{|S_{i,j}|^2}{r_{i,j}^2} \right)^{-\frac{2+v}{2}} \tag{32}$$

where $r_{i,j}$ is defined as (26). Using MM algorithm, the score function is derived as:

$$\varphi_{\mathrm{t}}(r_{i,j}) = \left(1 + \frac{2}{v} \right) \left(r_{i,j}^2 + \frac{2 \left| \hat{S}_{i,j} \right|^2}{v} \right)^{-1}. \tag{33}$$

Using (33) and AIP, $\hat{\mathbf{w}}_{i,j}$ is updated in a manner similar to GGD-ILRMA-based SBSS AEC. Analogously, the update rules of NMF matrices are derived using the recursive approximation technique [16], [24], [27]:

$$t_{i,k}^{\mathrm{t}} \leftarrow t_{i,k}^{\mathrm{t}} \left[\frac{\sum_{j'=1}^{j} \alpha_{\mathrm{t}}^{j-j'} \left| \hat{S}_{i,j'} \right|^2 c_{i,j}^{\mathrm{t}} r_{i,j'}^{-p} v_{k,j'}^{\mathrm{t}}}{\sum_{j'=1}^{j} \alpha_{\mathrm{t}}^{j-j'} r_{i,j'}^{-p} v_{k,j'}^{\mathrm{t}}} \right]^{\frac{p}{p+2}}, \tag{34}$$

$$v_{k,j}^{\mathrm{t}} \leftarrow v_{k,j}^{\mathrm{t}} \left[\frac{\sum_{i=1}^{I} \left| \hat{S}_{i,j} \right|^2 c_{i,j}^{\mathrm{t}} r_{i,j}^{-p} t_{i,k}^{\mathrm{t}}}{\sum_{i=1}^{I} r_{i,j}^{-p} t_{i,k}^{\mathrm{t}}} \right]^{\frac{p}{p+2}}, \tag{35}$$

where

$$c_{i,j}^{\mathrm{t}} = \left(\frac{v}{v+2} r_{i,j'}^2 + \frac{2}{v+2} \left| \hat{S}_{i,j'} \right|^2 \right)^{-1} \tag{36}$$

and $\alpha_{\mathrm{t}} \in (0, 1)$ is a forgetting factor for NMF variables.

D. Semi-Supervised NMF (SSNMF)

In AEC, little attention is paid to online ILRMA-based SBSS algorithms due to the challenges of implementing NMF in real-time [14]. Nonetheless, in AEC applications, it is often feasible to obtain a registered voice dataset of a certain length beforehand. Therefore, the SSNMF [17] technique can address this issue. In SSNMF, the basis matrix is pre-trained on the registered speech dataset before real-time NMF processing. During online processing, this basis matrix is fixed, and only the time-dependent activation matrix is updated. This approach alleviates the clustering problem typically associated with NMF, as there is only one source and one reference. The SSNMF technique offers a promising alternative to the previously mentioned recursive approximation technique.

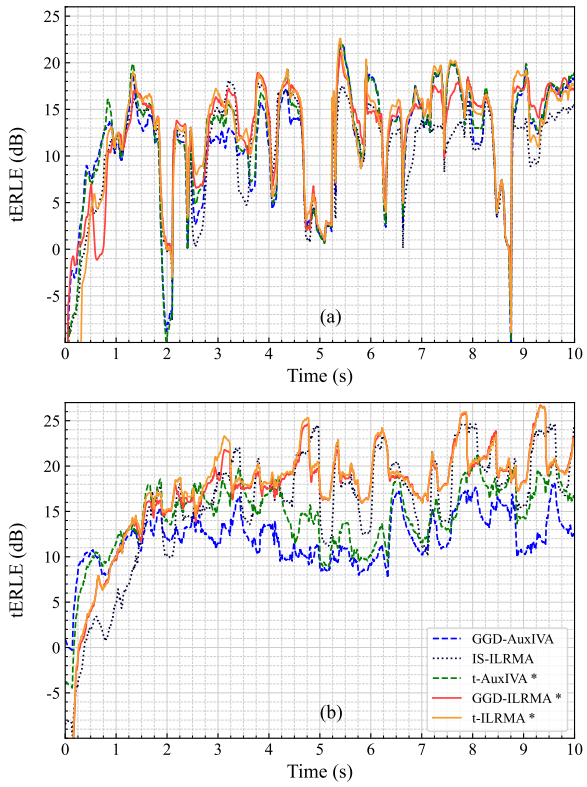

Fig. 1. Comparison of tERLE performance in different double-talk cases. (a) Speech (b) Music.

TABLE I
PARAMETER SETUP

Scenario	Speech				Music			
Algorithm	β	v	p	η	β	v	p	η
IS-ILRMA	-	-	-	0.992	-	-	-	0.990
GGD-AuxIVA	0.4	-	-	0.970	1	-	-	0.970
GGD-ILRMA	0.4	-	1	0.985	1	-	2	0.980
t-AuxIVA	-	10	-	0.970	-	30	-	0.970
t-ILRMA	-	10	1	0.985	-	30	2	0.980

IV. SIMULATIONS AND RESULTS

A. Experimental Setup

To maintain consistency, we use the same AIR setup as in [11], where the reverberation time, T_{60}, is approximately 300 ms. We consider the hard-clipping function [10], [19] to simulate loudspeaker distortions. We prepare two double-talk scenarios using speech and music signals, respectively. The signal-to-echo ratio (SER) is set to be 0 dB. In double-talk speech scenario, two 10-second-long male speech signals from the CMU ARCTIC dataset [28] are selected as the far-end and near-end signals, respectively. For SSNMF, we select random near-end registered speech signals to generate a 40-second pre-trained dataset. During the pre-training process, basis matrices are trained under 30 iterations with offline ILMRA algorithms. In double-talk music scenario, we use the same music signals as in the paper [10].

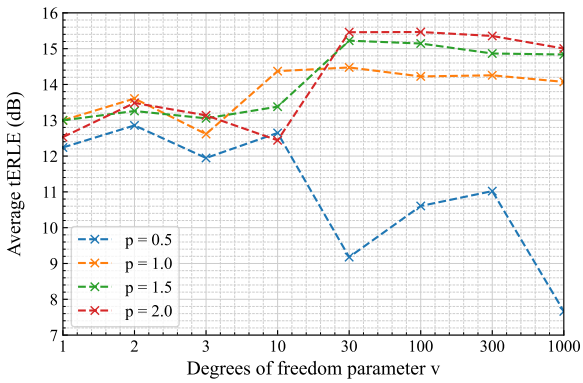

Fig. 2. Average SDR of 10-group double-talk Speech cases based on t-ILRMA-based SBSS-AEC

TABLE II
DOUBLE-TALK SPEECH SOUND QUALITY

Algorithms	PESQ	STOI
IS-ILRMA	1.502	0.921
GGD-AuxIVA	1.637	0.954
GGD-ILRMA (proposed)	1.679	0.953
t-AuxIVA (proposed)	1.635	0.955
t-ILRMA (proposed)	**1.792**	**0.960**

The sampling rate of all signals is 16 kHz. In short-time analysis, a Hann window with a length of 1024 samples and a 75% overlap between consecutive frames is used. The nonlinear series expansion order N and CTF filter length L are set to 3 and 5, respectively. The number of basis matrices K is set to be 8. The AIP initialization setups are consistent to that in [11]. The forgetting factors of the NMF basis matrix, α_{GGD} and α_{t}, are set below 0.02. The setup of the other parameters is shown in Table I.

In this section, we compare the AEC performance of proposed heavy-tailed distribution-based SBSS-AEC algorithms with IS-ILRMA-based and GGD-AuxIVA-based SBSS-AEC [10], [11]. We use true echo return loss enhancement (tERLE) [9] as performance metric. Additionally, perceptual evaluation of speech quality (PESQ) [29] and short time objective intelligibility (STOI) [30] are used as performance evaluation.

B. Simulation Results

First, we compare the performance in double-talk speech scenario. Fig. 1 (a) shows the tERLE performance of AuxIVA-based, SSNMF-ILRMA-based with GGD or Student's t-distribution, and IS-ILRMA-based online SBSS-AEC algorithms. It is evident that Student's t-distribution-based algorithms achieve comparable performance, and SSNMF-ILRMA-based methods outperform AuxIVA-based methods in the last nine seconds. All heavy-tailed distributions based algorithms perform better than IS-ILRMA based counterpart under optimal parameter settings, demonstrating robustness and superior

performance.

To illustrate the flexibility of heavy-tailed distribution-based algorithms more clearly, we present an example of parameter adjustment for t-ILRMA SBSS-AEC for 10-group double-talk speech cases from the CMU ARCTIC dataset. Figure 2 shows that when NMF domain parameter p is set to 2 and the degrees of freedom parameter v is set to 30, relevant algorithm can achieve best results. These curves highlight the process of modifying the source model. Table II shows the average PESQ and STOI performance of 10 double-talk speech cases with optimal parameter settings where proposed methods achieve better or comparable sound quality compared to baseline methods.

In double-talk music scenario, as displayed in Fig. 1 (b), it is quite obvious that ILRMA-based SBSS-AEC outperforms AuxIVA-based algorithms. Besides, all the proposed online ILRMA-based SBSS-AEC algorithms converge much faster than IS-ILRMA-based algorithms in the fisrt two seconds. This case demonstrates the superiority of ILRMA-based online SBSS-AEC algorithms in complex situations.

V. Conclusions

To further enhance the performance of SBSS-AEC algorithms, in this paper, we adopted two heavy-tailed distributions, i.e., the complex generalized super-Gaussian distribution (GGD) and Student's t-distribution. Based on them, we derived three novel algorithms, namely, GGD-ILRMA, t-AuxIVA and t-ILRMA based SBSS-AEC. Besides, online update rules for NMF were introduced. Simulations validated the effectiveness and superiority of proposed algorithms.

Acknowledgment

This work was supported by JSPS KAKENHI 23H03423.

References

[1] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in network and acoustic echo cancellation*. Springer, 2001.

[2] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1988, pp. 2570–2573.

[3] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: A practical approach*. John Wiley & Sons, 2005.

[4] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, no. 6, pp. 1140–1156, 2006.

[5] T. Gansler, S. L. Gay, M. M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Trans. Speech and Audio Process.*, vol. 8, no. 6, pp. 656–663, 2000.

[6] J. Gunther, "Learning echo paths during continuous double-talk using semi-blind source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 646–660, 2011.

[7] S. Makino, *Audio source separation*. Springer, 2018.

[8] 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, 2010.

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

[10] 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," *Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 88–95, 2023.

[11] X. Wang, Y. Yang, A. Brendel, T. Ueda, S. Makino, J. Benesty, *et al.*, "On semi-blind source separation-based approaches to nonlinear echo cancellation based on bilinear alternating optimization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2973–2987, 2024.

[12] 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 Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 756–760.

[13] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoust.*, 2011, pp. 189–192.

[14] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[15] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[16] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. Joint Workshop on Hands-free Speech Commun. and Mic. Arrays*, 2014, pp. 107–111.

[17] T. Wang, F. Yang, R. Zhu, and J. Yang, "Real-time independent vector analysis using semi-supervised nonnegative matrix factorization as a source model," in *Proc. Interspeech*, 2021, pp. 1842–1846.

[18] M. Schrammen, S. Kühl, S. Markovich-Golan, and P. Jax, "Efficient nonlinear acoustic echo cancellation by dual-stage multi-channel Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 975–979.

[19] 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.

[20] 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.

[21] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–4.

[22] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[23] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2006.

[24] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, *et al.*, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, pp. 1–25, 2018.

[25] Y. Liang, G. Chen, S. Naqvi, and J. A. Chambers, "Independent vector analysis with multivariate Student's t-distribution source prior for speech separation," *Electronics Letters*, vol. 49, no. 16, pp. 1035–1036, 2013.

[26] J. Harris, B. Rivet, S. M. Naqvi, J. A. Chambers, and C. Jutten, "Real-time independent vector analysis with Student's t source prior for convolutive speech mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 1856–1860.

[27] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, and N. Ono, "Independent low-rank matrix analysis based on complex Student's t-distribution for blind audio source separation," in *Proc. IEEE Int. Workshop Machine Learning for Sig. Process.*, 2017, pp. 1–6.

[28] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. ISCA Workshop on Speech Synthesis*, 2004, pp. 223–224.

[29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4214–4217.