A COMPUTATIONALLY EFFICIENT SEMI-BLIND SOURCE SEPARATION APPROACH FOR NONLINEAR ECHO CANCELLATION BASED ON AN ELEMENT-WISE ITERATIVE SOURCE STEERING

Kunxing Lu[⋆]*, Xianrui Wang*[⋆†]*, Tetsuya Ueda*[⋆]*, Shoji Makino*[⋆]*, and Jingdong Chen*[†]

[⋆] Waseda University, Japan
[†] Northwestern Polytechnical University, Xi'an, China

ABSTRACT

While the semi-blind source separation-based acoustic echo cancellation (SBSS-AEC) has received much research attention due to its promising performance during double-talk compared to the traditional adaptive algorithms, it suffers from system latency and nonlinear distortions. To circumvent these drawbacks, the recently developed ideas on convolutive transfer function (CTF) approximation and nonlinear expansion have been used in the iterative projection (IP)-based semi-blind source separation (SBSS) algorithm. However, because of the introduction of CTF approximation and nonlinear expansion, this algorithm becomes computationally very expensive, which makes it difficult to implement in embedded systems. Thus, we attempt in this paper to improve this IP-based algorithm, thereby developing an element-wise iterative source steering (EISS) algorithm. In comparison with the IP-based SBSS algorithm, the proposed algorithm is computationally much more efficient, especially when the nonlinear expansion order is high and the length of the CTF filter is long. Meanwhile, its AEC performance is as good as that of IP-based SBSS algorithm.

Index Terms— Semi-blind source separation, acoustic echo cancellation, convolutive transfer function approximation, nonlinear expansion, element-wise source steering

1. INTRODUCTION

In telecommunications or teleconferencing, acoustic echo, which is formed by the coupling between loudspeakers and microphones, is detrimental to full-duplex communication. One widely used way to eliminate the detrimental echo effects is through acoustic echo cancellation in which adaptive filters, such as the normalized least mean square (NLMS), recursive least mean square (RLS) and Kalman filters [1, 2], are used to identify the acoustic impulse response (AIR) [3]. While they have been widely used, adaptive filters generally suffer from great performance degradation or even divergence during double-talk in which both the far-end and near-end speech are present [4, 5].

One way to improve acoustic echo cancellation (AEC) performance during double-talk is through adopting the principle used in blind source separation (BSS) [6–9] to reformulate the AEC problem as one of semi-blind source separation (SBSS) [10, 11]. Several algorithms have been derived based on this SBSS framework, which have demonstrated promising performance [12–14]. However, those algorithms still suffer from a number of drawbacks. First, they are computationally very expensive, which makes them difficult to implement in embedded systems. A viable approach to reduce complexity is through using the so-called multiplicative transfer function model (MTF) [15] and performing echo cancellation in the short-time Fourier transform (STFT) domain [16, 17]. But this method increases the system latency as MTF requires the length of the analysis window to be longer than the effective part of AIR. Recently, the so-called convolutive transfer function (CTF) [18–20] model was applied to SBSS-AEC to achieve different compromise between the system latency and computational complexity [21, 22]. Second, the performance of SBSS-AEC suffers from significant degradation in the presence of loudspeaker nonlinearity, which happens often in small devices [23, 24]. One way to deal with this issue is through nonlinear AEC in which nonlinear expansion [25–27] is used to model the loudspeaker nonlinearity [28]. In [22], researchers proposed a framework which combines the CTF model and nonlinear expansion. An iterative projection (IP) [8] method is carried out to solve the corresponding optimization problem.

However, the IP-based algorithm still faces the challenge of computational complexity, as it requires to calculate the inverse of an auxiliary matrix. To circumvent this problem, we attempt in this work to reduce the complexity, thereby developing an element-wise iterative source steering (EISS) algorithm, which is an extension of the work in [29–31]. Since no matrix inverse is required, the developed algorithm is computationally much more efficient and its complexity is one order of magnitude lower than that of the IP-based algorithm, yet its performance is as good as that of the IP-based algorithm.

2. SIGNAL MODEL AND PROBLEM FORMULATION

Consider the full-duplex speech communication scenario where a microphone is used to pick up the sound signal from the near-end speaker and a loudspeaker is used to playback the signal from the far-end. The microphone output signal at time instant t, which is denoted as $y(t)$, can be written as

$$\begin{aligned} y(t) &= v(t) + s(t), \\ &= a(t) * f\left[x(t)\right] + s(t), \end{aligned} \tag{1}$$

where $s(t)$ denotes the near-end speech signal, $v(t) = a(t) * f\left[x(t)\right]$ is the nonlinear acoustic echo, $a(t)$ denotes the acoustic impulse response from the loudspeaker to the microphone, $*$ represents the linear convolution, $f(\cdot)$ stands for the response of the loudspeaker, which includes both the linear and nonlinear effects, and $x(t)$ is the

This work was supported by JSPS KAKENHI Grant Number 23H03423. The audio samples in this research are available at https://github.com/kunxinglu/audio_samples_ICASSP2024.

far-end signal, respectively. The problem of acoustic echo cancellation is to mitigate or eliminate the echo signal, i.e., $v(t)$, while preserving the near-end signal $s(t)$.

While they have been widely used in practical systems, adaptive filtering algorithms often suffer from great performance degradation in the presence of loundspeaker nonlinearity. One way to deal with loundspeaker nonlinear distortion is to approximate $f\left[x(t)\right]$ through a Pth-order basis-generic expansion [26], i.e.,

$$f\left[x(t)\right] = \sum_{p=0}^{P-1} c_p \phi_p\left[x\left(t\right)\right], \quad p = 0, 1, \ldots, P-1, \tag{2}$$

where $\phi_p\left(\cdot\right)$ is the p-th order basis function and c_p is the corresponding coefficient. For real-valued signal, the following expansion can be used,

$$\phi_p\left[x\left(t\right)\right] = x^{2p+1}(t). \tag{3}$$

Substituting (2) into (1) gives

$$\begin{aligned} y(t) &= \sum_{p=0}^{P-1} c_p a(t) * \phi_p\left[x\left(t\right)\right] + s(t), \\ &= \sum_{p=0}^{P-1} a'_p(t) * \phi_p\left[x(t)\right] + s(t), \end{aligned} \tag{4}$$

where $a'_p(t) = c_p a(t)$ denote the echo path of the p-th order expanded signal $\phi_p\left[x(t)\right]$.

To reduce the computational complexity, the MTF model is introduced, which requires the analysis window to be longer than the effective part of AIR, leading to larger system latency. To achieve proper compromise between the computational complexity and system latency, the CTF approximation is subsequently adopted. The signal model in (4) is then written in the STFT domain as

$$Y_{i,j} = \sum_{p=0}^{P-1} \sum_{l=0}^{L-1} A'_{p,i,l} X_{\phi,p,i,j-l} + S_{i,j}, \tag{5}$$

where i and j denote the frequency and time-frame indexes, L is the length of the CTF filter, and $Y_{i,j}$, $A'_{p,i,j}$, $X_{\phi,p,i,j}$, $S_{i,j}$ denote, respectively, the STFTs of $y(t)$, $a'_p(t)$, $\phi_p\left[x(t)\right]$ and $s(t)$. Putting (5) into a vector/matrix form gives

$$\tilde{\mathbf{y}}_{i,j} = \tilde{\mathbf{H}}_{i,j} \tilde{\mathbf{s}}_{i,j}, \tag{6}$$

where

$$\tilde{\mathbf{y}}_{i,j} = \begin{bmatrix} Y_{i,j} & \mathbf{x}_{0,i,j}^T & \cdots & \mathbf{x}_{P-1,i,j}^T \end{bmatrix}^T, \tag{7}$$

$$Y_{i,j} = \sum_{p=0}^{P-1} \mathbf{a}_{p,i,j}^T \mathbf{x}_{p,i,j} + S_{i,j}, \tag{8}$$

$$\mathbf{a}_{p,i,j} = \begin{bmatrix} A'_{p,i,0} & \cdots & A'_{p,i,L-1} \end{bmatrix}^T, \tag{9}$$

$$\mathbf{x}_{p,i,j} = \begin{bmatrix} X_{\phi,p,i,j} & \cdots & X_{\phi,p,i,j-L+1} \end{bmatrix}^T, \tag{10}$$

$$\tilde{\mathbf{s}}_{i,j} = \begin{bmatrix} S_{i,j} & \mathbf{x}_{0,i,j}^T & \cdots & \mathbf{x}_{P-1,i,j}^T \end{bmatrix}^T, \tag{11}$$

$$\tilde{\mathbf{H}}_{i,j} = \begin{bmatrix} 1 & \mathbf{a}_{i,j}^T \\ \mathbf{0}_{PL\times 1} & \mathbf{I}_{PL} \end{bmatrix}, \tag{12}$$

$$\mathbf{a}_{i,j} = \begin{bmatrix} \mathbf{a}_{0,i,j}^T & \mathbf{a}_{1,i,j}^T & \ldots & \mathbf{a}_{P-1,i,j}^T \end{bmatrix}^T, \tag{13}$$

the superscript T denotes the transpose operation, and $\mathbf{I}_{PL}$ denotes the identity matrix of size $PL \times PL$. Note that the $\tilde{\mathbf{H}}_{i,j}$ matrix is of size $(PL+1) \times (PL+1)$, which is called the mixing matrix in the literature of BSS, $\mathbf{0}_{PL\times 1}$ is a zero vector of length PL.

Following the notation in BSS and echo cancellation, we now define the demixing matrix $\tilde{\mathbf{W}}_{i,j}$ as

$$\tilde{\mathbf{W}}_{i,j} = \begin{bmatrix} 1 & \mathbf{b}_{i,j}^T \\ \mathbf{0}_{PL\times 1} & \mathbf{I}_{PL} \end{bmatrix}, \tag{14}$$

where $\mathbf{b}_{i,j}$ is a column vector with PL parameters to be estimated. The near-end signal extraction filter can then be expressed as $\tilde{\mathbf{w}}_{i,j}^H = [1 \quad \mathbf{b}_{i,j}^T]$. Applying the near-end signal extraction filter to the input signal gives the near-end signal, i.e.,

$$\hat{S}_{i,j} = \tilde{\mathbf{w}}_{i,j}^H \tilde{\mathbf{y}}_{i,j}. \tag{15}$$

Given the aforementioned signal model and problem formulation, the objective of nonlinear SBSS-AEC is to estimate $\tilde{\mathbf{w}}_{i,j}$ by exploiting independence between the near-end and the reference signals.

3. NONLINEAR SBSS-AEC ALGORITHMS

3.1. Probabilistic Model

We consider to model the source signal with a generalized Gaussian distribution, i.e.,

$$p\left(\mathbf{s}_j\right) \propto \exp\left[-\left(\frac{\left\|\mathbf{s}_j\right\|_2}{\gamma}\right)^{\beta}\right], \tag{16}$$

where

$$\mathbf{s}_j = \left[S_{1,j} \quad S_{2,j} \quad \ldots \quad S_{I,j}\right]^T, \tag{17}$$

$\left\|\cdot\right\|_2$ stands for ℓ_2 norm. γ and β are the scale and shape parameters, respectively. Since the reference signal is accessible, the negative log-likelihood function can then be calculated as

$$\begin{aligned} \mathcal{L}_j = &- \frac{1}{\sum_{j'=1}^{j} \alpha^{j-j'}} \sum_{j'=1}^{j} \alpha^{j-j'} \log p\left(\mathbf{s}_{j'}\right) \\ &- 2\sum_{i=1}^{I} \log\left|\det \tilde{\mathbf{W}}_{i,j}\right|, \end{aligned} \tag{18}$$

where $\alpha \in (0,1)$ is a forgetting factor. By using the well known majorization-minimization (MM) method [32], the following auxiliary function can be obtained

$$\mathcal{L}_j^+ = \sum_{i=1}^{I} \tilde{\mathbf{w}}_{i,j}^H \mathbf{V}_{i,j} \tilde{\mathbf{w}}_{i,j} - 2\sum_{i=1}^{I} \log\left|\det \tilde{\mathbf{W}}_{i,j}\right|. \tag{19}$$

To track time-varying signals and acoustic environments, the recursive estimation of $\mathbf{V}_{i,j}$ is generally used, i.e.,

$$\mathbf{V}_{i,j} = \alpha \mathbf{V}_{i,j-1} + (1-\alpha)\varphi(r_j)\tilde{\mathbf{y}}_{i,j}\tilde{\mathbf{y}}_{i,j}^H, \tag{20}$$

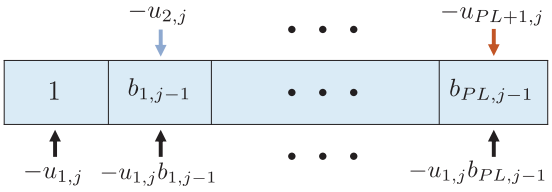

Fig. 1. Illustration of EISS update rule.

and

$$\varphi(r_j) = r_j^{\beta-2}, \tag{21}$$

$$r_j = \sqrt{\textstyle\sum_{i=1}^{I} |\tilde{\mathbf{w}}_{i,j-1}^H \tilde{\mathbf{y}}_{i,j}|^2}. \tag{22}$$

Note that the norm of the near-end signal extraction filter $\tilde{\mathbf{w}}_{i,j}$ does not affect the independence criterion. Therefore, a two-stage estimation strategy can be used in which the extraction filter is updated in the first stage using the maximum likelihood criterion and the updated filter is then normalized in the second stage such that its first element is equal to 1. Since the extraction filters at different frequency bins are estimated independently, we shall omit the frequency index i in the rest parts of this paper without introducing any confusion.

3.2. Conventional Iterative Projection Based Method

The IP method can be derived by identifying the Wirtinger derivative of (19) with respect to $\tilde{\mathbf{w}}_{i,j}$ and then forcing the result equal to 0. The update rules are shown as follows:

$$\tilde{\mathbf{w}}_j \leftarrow (\tilde{\mathbf{W}}_{j-1}\mathbf{V}_j)^{-1}\mathbf{e}_1 = \mathbf{V}_j^{-1}\mathbf{e}_1, \tag{23}$$

$$\tilde{\mathbf{w}}_j \leftarrow \frac{\tilde{\mathbf{w}}_j}{w_{1,j}}, \tag{24}$$

where $\mathbf{e}_1$ is the first column of $\mathbf{I}_{PL+1}$ and $w_{1,j}$ is the first element of $\tilde{\mathbf{w}}_j$.

3.3. Proposed Element-wise Iterative Source Steering Method

Implementation of the IP method requires to compute the inverse of the $\mathbf{V}_j$ matrix every frame, which makes the algorithm computationally very expensive. To reduce the complexity, we propose to update the near-end signal extraction filter with the EISS method in the first stage [29–31], i.e.,

$$\begin{cases} w_{1,j} \leftarrow w_{1,j-1} - u_{1,j}, & \text{if } k = 1 \\ w_{k,j} \leftarrow w_{k,j-1} - u_{1,j}w_{k,j-1} - u_{k,j}, & \text{if } k \neq 1 \end{cases} \tag{25}$$

where $w_{k,j}$, $k = 1, 2, \ldots, PL+1$, is k-th element of $\tilde{\mathbf{w}}_j$ and $u_{k,j}$ is a parameter to estimate. All the $u_{k,j}$'s need to be estimated sequentially. In other words, the algorithm first computes $u_{1,j}$ to update all the elements related to w_{j-1}; it then computes $u_{2,j}$ to update $b_{1,j-1}$; it subsequently computes $u_{3,j}$ to update $b_{2,j-1}$, and so forth. The EISS update rules are illustrated in Fig. 1.

Substituting (25) into the auxiliary function, we obtain

$$\mathcal{L}_j^+(u_{k,j}) = -2\log|1 - u_{1,j}| + \mathbf{d}_j^H \mathbf{V}_j \mathbf{d}_j, \tag{26}$$

where

$$\begin{aligned} \mathbf{d}_j^H = &\tilde{\mathbf{w}}_{j-1}^H - [u_{1,j} \quad u_{1,j}b_{1,j-1} + u_{2,j} \\ &\cdots u_{1,j}b_{PL,j-1} + u_{PL+1,j}]. \end{aligned} \tag{27}$$

Identifying the Wirtinger derivative of $\mathcal{L}_j^+(u_{k,j})$ with respect to $u_{k,j}^*$ and forcing the result to be 0, one can obtain the following solution

$$u_{k,j} = \begin{cases} \dfrac{\tilde{\mathbf{w}}_{j-1}^H \mathbf{v}_{k,j}}{V_j(k,k)}, & k \neq 1 \\ 1 - (\tilde{\mathbf{w}}_{j-1}^H \mathbf{V}_j \tilde{\mathbf{w}}_{j-1})^{-\frac{1}{2}}, & k = 1 \end{cases}, \tag{28}$$

where $\mathbf{v}_k$ denotes the k-th column of $\mathbf{V}_j$ and $V_j(k,k)$ stands for the (k,k)-th element of $\mathbf{V}_j$.

4. COMPLEXITY ANALYSIS

In this section, we present the complexity of the IP and EISS algorithms in terms of the number of multiplications/divisions needed per frequency bin and time frame. For the IP algorithm, the complexity is dominated by computing the inverse of the covariance matrix $\mathbf{V}_j$, which has a complexity proportional to $\mathcal{O}(PL+1)^3$. The complexity for all the other computations is of $\mathcal{O}(PL+1)$. For regular setup, $\mathcal{O}(PL+1)$ is much smaller than $\mathcal{O}(PL+1)^3$. Therefore, the complexity of the IP method is

$$\mathcal{C}_{\text{IP}} \propto \mathcal{O}\left[(PL)^3\right]. \tag{29}$$

The EISS algorithm needs to estimate $PL+1$ coefficients. Computation of $u_{k,j}, k \neq 1$ has a complexity of $\mathcal{O}\left[PL(PL+1)\right]$. The complexity for computing $u_{1,j}$ is $\mathcal{O}\left[(PL+1)^2\right]$ and it is $O(PL+1)$ for all the other operations. Similarly, if we only consider the operations that dominate the complexity, the overall complexity of EISS method is

$$\mathcal{C}_{\text{EISS}} \propto \mathcal{O}\left[(PL)^2\right], \tag{30}$$

which is one order lower than that of the IP method.

5. SIMULATIONS AND EXPERIMENTS

5.1. Experimental Setup

In this section, the proposed EISS-based algorithm is compared with the IP-based algorithm and the single-microphone form of the state-space model (SSM)-based nonlinear acoustic echo cancellation algorithm proposed in [27]. We evaluated the proposed AEC algorithms with the help of objective measures to quantify the performance in terms of echo reduction and speech distortion. For the single-talk case, echo return loss enhancement (ERLE) [26] is used as the performance metric, and for double-talk, true ERLE (tERLE) [17] is used. Besides, perceptual evaluation of speech quality (PESQ) [33], and short time objective intelligibility (STOI) [34] are also used for performance evaluation. The sampling rate for all the signals in this work is 16 kHz.

To assess the efficiency of our algorithm, we also conducted a comparative analysis of the runtime performance between our algorithm and an IP-based algorithm. We executed 100 signals, each lasting 10 seconds, on a laptop equipped with an i7-10750H CPU and computed the average runtime of each signal as the final test result. The average runtime is showned in Fig. 3. For the short-time analysis, the frame length is 1024-point long with an overlap factor

of 75%. The Hanning window is applied and the windowed signal is then transformed into the STFT domain with a 1024-point fast Fourier transform (FFT). To balance the computational complexity and performance, the nonlinear expansion order P is set to 3 and the length of CTF filter L is set to 5. The forgetting factor α is set to 0.992. The shape parameter β is set to 0.4. In all experiments, the demixing matrix $\tilde{\mathbf{W}}$ is initialized as an identity matrix $\mathbf{I}$ and the auxiliary matrix $\mathbf{V}$ is initialized as $10^{-3} \times \mathbf{I}$.

5.2. AEC Performance for Hard Clipping Mapping

In this experiment, we validate the ability of EISS to handle nonlinear distortion in both single-talk and double-talk scenarios. We use the same data as in [22]. The hard clipping function [26] is used to simulate the loudspeaker nonlinearity, in which the clipping threshold is set to $0.2max|x(t)|$. The reverberation time T_{60} is approximately 300 ms. The signal-to-echo ratio (SER) for double-talk is 0 dB. Figure 2 (a) and (b) show the performance of the IP and EISS methods in the double-talk and single-talk situations, respectively. One can see that the ERLE and tERLE of EISS and IP algorithms are almost the same, but when compared to SSM, both of them have significantly higher values. Therefore, the proposed algorithm demonstrates superior AEC performance compared to the conventional SSM algorithm.

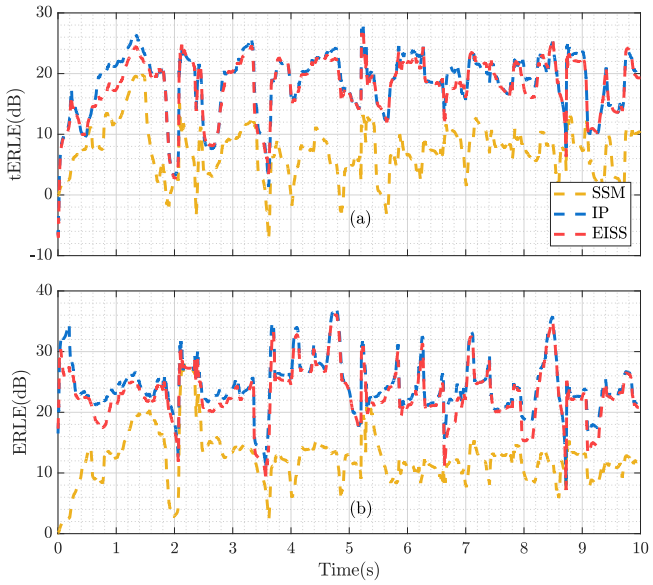

Fig. 2. The performance of SSM, IP and EISS: (a). tERLE in double-talk situation, (b). ERLE in single-talk situation.

5.3. Overall Performance on the AEC Challenge Dataset

In this experiment, a total of 30 signals are arbitrarily taken from the AEC challenge synthetic dataset [35]. The nonlinear far-end signals are generated by clipping the maximum amplitude or by applying the sigmoidal function [36] and learned distortion functions to far-end signal. The SER of these signals ranges from -10 dB to 10 dB and the T_{60} ranges from 200 ms to 1200 ms. Table 1 lists the overall performance of the three algorithms in terms of PESQ, STOI, and tERLE. It can be seen that the proposed algorithm significantly outperforms traditional nonlinear AEC algorithms across various noise and reverberation environments and exhibits similar performance to the IP-based algorithm.

Table 1. Performance of SSM, IP and EISS.

Algorithm	PESQ	STOI	tERLE
SSM	1.57	0.87	8.77
IP	1.89	0.93	12.89
EISS	1.9	0.94	12.63

5.4. Runtime Comparison

In the last set of experiments, we compare the runtime of the IP and EISS method with the same setup as described previously. The time measured here includes the time to compute and update the covariance matrix and auxiliary variables. In this experiment, the nonlinear expansion order P is set to 3 and 4, respectively. The CTF filter length, i.e., L, varies from 2 to 12. As shown in Fig. 3, the EISS

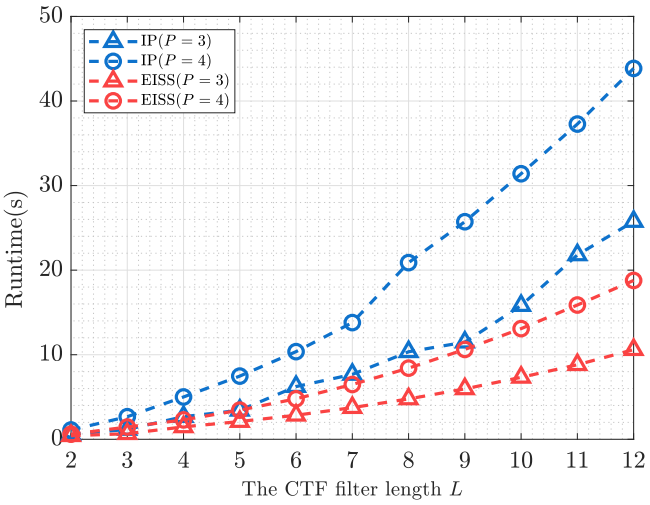

Fig. 3. The runtime comparison between IP and EISS.

method is computationally much more efficient than the IP method and the difference becomes much more dramatic as values of P and L increases.

6. CONCLUSION

This paper investigated the problem of AEC in the presence of doubletalk and loudspeaker nonlinearity within reverberant environment. Following the framework of SBSS, we developed an element-wise source steering algorithm, which combines the CTF model and nonlinear expansion into the SBSS framework. Unlike the conventional IP-based SBSS method, which requires to compute the inverse of the auxiliary matrix, the proposed algorithm applies an element-wise update strategy, in which no matrix inverse is involved, and as a result, its computational complexity is one order of magnitude lower than that of the IP-based algorithm. Moreover, experiments showed that this efficient algorithm is able to achieve similar performance to that of the IP-based algorithm.

7. REFERENCES

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

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

[3] X. Wang, G. Huang, J. Benesty, J. Chen, and I. Cohen, "Time difference of arrival estimation based on a Kronecker product decomposition," *IEEE Signal Process. Lett.*, vol. 28, pp. 51–55, 2020.

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

[5] H. Buchner, J. Benesty, T. Gansler, and W. Kellermann, "Robust extended multidelay filter and double-talk detector for acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1633–1644, 2006.

[6] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[7] 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.

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

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

[10] M. Joho, H. Mathis, and G. S. Moschytz, "Combined blind/nonblind source separation based on the natural gradient," *IEEE Signal Process. Lett.*, vol. 8, no. 8, pp. 236–238, 2001.

[11] S. Miyabe, T. Takatani, H. Saruwatari, K. Shikano, and Y. Tatekura, "Barge-in-and noise-free spoken dialogue interface based on sound field control and semi-blind source separation," in *Proc. EUSIPCO*, 2007, pp. 232–236.

[12] 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.

[13] Z. Koldovskỳ, J. Málek, M. Müller, and P. Tichavskjỳ, "On semi-blind estimation of echo paths during double-talk based on nonstationarity," in *Proc. IWAENC*, 2014, pp. 198–202.

[14] J. Gunther and T. Moon, "Blind acoustic echo cancellation without double-talk detection," in *Proc. WASPAA*, 2015, pp. 1–5.

[15] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, 2007.

[16] T. S. Wada, S. Miyabe, and B.-H. F. Juang, "Use of decorrelation procedure for source and echo suppression," in *Proc. IWAENC*, 2008, pp. 1–5.

[17] 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.

[18] 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.

[19] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1420–1434, 2009.

[20] 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.

[21] Z. 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.

[22] 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.

[23] R. Niemistö and T. Mäkelä, "On performance of linear adaptive filtering algorithms in acoustic echo control in presence of distorting loudspeakers," in *Proc. IWAENC*, 2003, pp. 79–82.

[24] M. I. Mossi, N. W. Evans, and C. Beaugeant, "An assessment of linear adaptive filter performance with nonlinear distortions," in *Proc. IEEE ICASSP*, 2010, pp. 313–316.

[25] S. Malik and G. Enzner, "Fourier expansion of Hammerstein models for nonlinear acoustic system identification," in *Proc. IEEE ICASSP*, 2011, pp. 85–88.

[26] 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.

[27] J. Park and J.-H. Chang, "State-space microphone array nonlinear acoustic echo cancellation using multi-microphone near-end speech covariance," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1520–1534, 2019.

[28] 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, 2021.

[29] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE ICASSP*, 2020, pp. 236–240.

[30] T. Nakashima and N. Ono, "Inverse-free online independent vector analysis with flexible iterative source steering," in *Proc. APSIPA*, 2022, pp. 749–753.

[31] T. Nakashima, R. Ikeshita, N. Ono, S. Araki, and T. Nakatani, "Fast online source steering algorithm for tracking single moving source using online independent vector analysis," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[32] K. Lange, *MM optimization algorithms*. SIAM, 2016.

[33] 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 ICASSP*, vol. 2, 2001, pp. 749–752.

[34] 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 ICASSP*, 2010, pp. 4214–4217.

[35] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sørensen, and R. Aichner, "ICASSP 2022 acoustic echo cancellation challenge," in *Proc. IEEE ICASSP*, 2022, pp. 9107–9111.

[36] C. M. Lee, J. W. Shin, and N. S. Kim, "DNN-based residual echo suppression," in *Proc. Interspeech*, 2015, pp. 1175–1179.