Enhancing the Element-wise Source Steering Algorithm Using Crossband Filters for Double-talk Robust Acoustic Echo Cancellation

Chengxiao Zhao, Kaien Mo, Liyuan Zhang and Shoji Makino

Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan
E-mail: {zhaochengxiao@akane, kaien.m@ruri, ly.zhang@akane, s.makino@}.waseda.jp

Abstract

Semi-blind source separation-based acoustic echo cancelation (SBSS-AEC) methods have gained significant attention for their exceptional ability to effectively manage both single-talk and double-talk situations. The convolutive transfer function (CTF) approximation with band-to-band filters is often adopted in such methods to make a trade-off between the computational complexity and system latency. However, CTF neglects occasions where the power of a single time-frequency frame leaks into neighboring frequency bins, especially in highly reverberant conditions. In this paper, we propose an extended version of the SBSS-AEC method to enhance modeling accuracy and robustness by introducing cross-band filters into the modeling of the source extraction filter. To manage the increased complexity introduced by the cross-band coefficients, we utilize the recently proposed element-wise iterative source steering (EISS) algorithm as a more efficient alternative to the iterative projection (IP) algorithm. Simulations demonstrate the effectiveness of our proposal.

1. Introduction

In modern teleconferencing systems, acoustic echo arises due to the loudspeaker-microphone coupling and severely degrades the quality of communication. Consequently, numerous acoustic echo cancellation (AEC) techniques have been developed to remove the echo. A commonly adopted method is to reformulate the AEC problem as a semi-blind source separation (SBSS) problem. Such methods, referred to as SBSS-AEC, showcase stable performance in both double-talk and single-talk situations, which depend on the status of far-end and near-end speakers. SBSS is an extension of blind source separation (BSS) [1] where partial information of the source signals, i.e., reference signals, is known beforehand. In real-world applications, assuming the known far-end signal as the reference signal is typical, and implementing SBSS-AEC algorithms is typical.

The latest SBSS-AEC proposals [2, 6–8] are derived in the short-time Fourier transform (STFT) domain, among which the convolutive transfer function (CTF) approximation is commonly utilized. The CTF introduces a convolutional filter for source separation in the STFT domain. This filter modeling alleviates the constraint between the STFT frame length and the reverberation time, enabling CTF-based methods to capture inter-frame information and achieve better performance in highly reverberant environments [2]. However, the CTF model treats the mixing process as independent across frequency bins, which is an incomplete representation as demonstrated in [3]. The power of a signal time-frequency frame will leak into adjacent frames and neighboring frequency bins, especially when the reverberation is high. Specifically, the convolution in a linear-time-invariant system involves all clean subbands and their associated cross-band filters, which capture the dependencies and interactions between different frequency subbands. Therefore, considering cross-band filters is reasonably necessary.

In this paper, we aim to enhance the performance of the SBSS-AEC algorithm by incorporating cross-band filters [4]. However, introducing the cross-band coefficients will significantly increase the filter length, and the inherent complexity of the original iterative projection (IP) algorithm [5] in conventional methods renders it impractical for real-world applications. To address this, we adopt our recently proposed element-wise iterative source steering (EISS) algorithm [6,7]. The EISS algorithm is computationally much more efficient since it avoids the matrix inverse in IP. To validate our proposal, several simulations were conducted and the results confirmed the effectiveness of the proposed approach.

2. Signal model and problem fromulation

In a double-talk communication system, the purpose of a microphone is to capture the speaker's voice from near-end. A loudspeaker is used to broadcast the sound from the far end. During the broadcast process, the far-end signal is convolved with acoustic impulse response (AIR), which is inherent in the room. The echo is generated in this situation. Finally, the microphone received a mixed signal of near-end signal and echo signal. In the time domain at time index t, the process can be expressed as

$$\begin{aligned} y(t) &= v(t) + s(t), \\ &= a(t) \star x(t) + s(t), \end{aligned} \tag{1}$$

where $\star$ denotes convolution operation, $y(t)$, $x(t)$, $s(t)$, $v(t)$ and $a(t)$ are microphone observation signal, far-end signal, near-end signal, echo signal, and AIR, respectively.

The AIR is generally very long in a reverberant environment, and it is almost impossible to identify it in the time domain due to the heavy complexity burden. A proper alternative is to implement AEC in the STFT domain at the expense of increasing system latency. With the exact CTF model, the microphone signal can be approximated as

$$Y_{i,j} = \sum_{p=-P}^{P} \sum_{l=0}^{L-1} A_{i+p,j,l} X_{i+p,j-l} + S_{i,j}, \tag{2}$$

where i and j are the frequency and time frame indexes, P is related to the number of cross-band filters we used., L is the length of the band-to-band filter, $A_{i+p,j,l}$ represents the filter coefficient, and $X_{i,j}$, $Y_{i,j}$, $S_{i,j}$ denote, respectively, the STFTs of $x(t)$, $y(t)$ and $s(t)$. Note that the near-end speech is generally close to the microphone, and we do not apply dereverberation to it, so we use the STFT representation of its microphone image in the above equation.

From the perspective of SBSS-AEC, the mixing process in (2) can be rewritten as $\tilde{\mathbf{y}}_{i,j} = \mathbf{A}_{i,j}\tilde{\mathbf{s}}_{i,j}$, where

$$\mathbf{A}_{i,j} = \begin{bmatrix} 1 & \tilde{\mathbf{a}}_{i,j}^T \\ \mathbf{0}_{\tilde{L}\times 1} & \mathbf{I}_{\tilde{L}} \end{bmatrix}, \tag{3}$$

is the mixing matrix, in which

$$\tilde{\mathbf{a}}_{i,j}^T = \begin{bmatrix} \mathbf{a}_{i-P,j}^T & \mathbf{a}_{i-P+1,j}^T & \cdots & \mathbf{a}_{i+P,j}^T \end{bmatrix}^T, \tag{4}$$

$$\mathbf{a}_{i-p,j}^T = \begin{bmatrix} A_{i-p,j,0} & A_{i-p,j,1} & \cdots & A_{i-p,j,L-1} \end{bmatrix}^T. \tag{5}$$

The superscript T denotes the transpose operation, $\tilde{L} = (2P+1)L$, $\boldsymbol{A}_{i,j}$ is the mixing matrix with the size of $(\tilde{L}+1) \times (\tilde{L}+1)$, $\mathbf{0}_{\tilde{L}\times 1}$ is a column vector of length $\tilde{L}$ with all elements equal to 0, and $\boldsymbol{I}_{\tilde{L}}$ is an identity matrix of size $\tilde{L} \times \tilde{L}$. The concatenated signal vectors $\tilde{\mathbf{y}}_{i,j}$ and $\tilde{\mathbf{s}}_{i,j}$ are defined as

$$\tilde{\mathbf{y}}_{i,j} = \begin{bmatrix} Y_{i,j} & \mathbf{x}_{i-P,j}^T & \mathbf{x}_{i-P+1,j}^T & \cdots & \mathbf{x}_{i+P,j}^T \end{bmatrix}^T, \tag{6}$$

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

and

$$\mathbf{x}_{i,j} = \begin{bmatrix} X_{i,j} & X_{i,j-1} & \cdots & X_{i,j-L+1} \end{bmatrix}^T. \tag{8}$$

The target of AEC is to remove the echo signal and extract the clean near-end signal from the microphone output signal. The far-end signal $\mathbf{x}_{i,j}$ can be regarded as the known reference signal. It is obvious that $\mathbf{A}_{i,j}$ is non-singular. We define its inverse as the demixing matrix, i.e. $\mathbf{W}_{i,j} = \mathbf{A}_{i,j}^{-1}$. The close form of $\mathbf{W}_{i,j}$ is defined as

$$\mathbf{W}_{i,j} = \begin{bmatrix} 1 & -\tilde{\mathbf{a}}_{i,j}^T \\ \mathbf{0}_{\tilde{L}\times 1} & \mathbf{I}_{\tilde{L}} \end{bmatrix}. \tag{9}$$

The first row of the demixing matrix is defined as the near-end source extraction filter, i.e.,

$$\mathbf{w}_{i,j}^H = \begin{bmatrix} 1 & -\tilde{\mathbf{a}}_{i,j}^T \end{bmatrix}, \tag{10}$$

where $(\cdot)^H$ denotes conjugate transpose, $-\tilde{\mathbf{a}}_{i,j}$ is a column vector of length $\tilde{L}$. Now we have

$$\hat{S}_{i,j} = \mathbf{w}_{i,j}^H \tilde{\mathbf{y}}_{i,j}. \tag{11}$$

Now, the problem is formulated as an adaptive estimate of $\mathbf{w}_{i,j}$ by adapting the independence between the near-end signal and the far-end signal, $\hat{S}_{i,j}$ is the estimated signal.

3. SBSS-AEC algorithms with crossband filtering

The near-end signal is assumed as following a generalized Gaussian distribution [9], i.e.,

$$p(\mathbf{s}_j) \propto \exp\left[-\left(\frac{\|\mathbf{s}_j\|_2}{\gamma}\right)^{\beta}\right], \tag{12}$$

where

$$\mathbf{s}_j = \begin{bmatrix} S_{1,j} & S_{2,j} & \ldots & S_{I,j} \end{bmatrix}^T, \tag{13}$$

where $\|\cdot\|_2$ denotes ℓ_2 norm, and $\gamma > 0$, $0 < \beta \leq 2$ are two shape paramters. By exploiting the mutual independence between far-end and near-end signals, the negative log-likelihood function can be derived as

$$\begin{aligned} \mathcal{L} = &-\frac{1}{\sum_{j'=1}^{j} \alpha^{j-j'}} \sum_{j'=1}^{j} \alpha^{j-j'} \log p(s_{j'}) \\ &- 2\sum_{i=1}^{I} \log|\det \mathbf{W}_{i,j}|, \end{aligned} \tag{14}$$

where $0 < \alpha < 1$ is a forgetting factor. Utilizing the majorization-minimization (MM) optimization method [10], the following cost function is adopted as the optimization criteria

$$\mathcal{L}^{+} = \sum_{i=1}^{I} \mathbf{w}_{i,j}^H \mathbf{V}_{i,j} \mathbf{w}_{i,j} - 2\sum_{i=1}^{I} \log|\det \mathbf{W}_{i,j}|, \tag{15}$$

where the auxiliary matrix $\mathbf{V}_{i,j}$ is updated recursively, i.e.,

$$\mathbf{V}_{i,j} = \alpha \mathbf{V}_{i,j-1} + (1-\alpha)\varphi(r_j)\tilde{\mathbf{y}}_{i,j}\tilde{\mathbf{y}}_{i,j}^H, \tag{16}$$

the weight is calculated from the perspective of the maximum likelihood criterion, i.e.,

$$\varphi(r_j) = |\mathbf{w}_{i,j-1}^H \tilde{\mathbf{y}}_{i,j}|^{\beta-2}. \tag{17}$$

As the filter is of $(2P+1)L+1$ parameters, the IP algorithm [5] is impractical for the real-time system as the matrix inverse is involved. Therefore, we adopt the EISS algorithm to update instead of IP [6, 7]. EISS is a computationally efficient method for decreasing (15). Without calculating the inverse of the auxiliary matrix $\mathbf{V}_{i,j}$, it updates each element in $\mathbf{w}_{i,j}$ individually with following update rule

$$\begin{cases} w_{1,i,j} \leftarrow w_{1,i,j-1} - u_{1,i,j}, & k=1 \\ w_{k,i,j} \leftarrow (1-u_{1,i,j})w_{k,i,j-1} - u_{k,i,j}, & k=2,...,\tilde{L} \end{cases} \tag{18}$$

where $w_{k,i,j}$ is the k-th element of $\mathbf{w}_{i,j}$ and $u_{k,i,j}$ is the steering step size. Substituting (18) into the auxiliary function (15), we have

$$\begin{aligned} \mathcal{L}^+ =& -2\log|1-u_{1,i,j}| \\ &+ (\mathbf{w}_{i,j-1} - \mathbf{u}_{i,j})^H \mathbf{V}_{i,j} (\mathbf{w}_{i,j-1} - \mathbf{u}_{i,j}), \end{aligned} \tag{19}$$

In EISS, the algorithm first estimates $u_{1,i,j}$ and updates all the elements in $w_{i,j-1}$. Then, it estimates $u_{2,i,j}$ and updates $-\tilde{a}_{1,i,j-1}$. In this sequence, it finally estimates $u_{\tilde{L}+1,i,j}$ to updates $-\tilde{a}_{\tilde{L},i,j-1}$. The process can be formulated as

$$\begin{aligned} \mathbf{u}_{i,j} = \Big[& u_{1,i,j} \quad -u_{1,i,j}\tilde{a}_{1,i,j-1} + u_{2,i,j} \\ & \cdots \quad -u_{1,i,j}\tilde{a}_{\tilde{L},i,j-1} + u_{\tilde{L}+1,i,j} \Big]. \end{aligned} \tag{20}$$

Calculating the derivative of $\mathcal{L}^+$ with respect to $(u_{k,i,j})^*$ and setting the result to 0, we can obtain

$$u_{k,i,j} = \begin{cases} 1 - \left(\mathbf{w}_{i,j-1}^H \mathbf{V}_{i,j} \mathbf{w}_{i,j-1}\right)^{-\frac{1}{2}}, & k=1 \\ \dfrac{\mathbf{w}_{i,j-1}^H \mathbf{v}_{k,i,j}}{V_{i,j}(k,k)}, & k=2,\ldots,\tilde{L} \end{cases} \tag{21}$$

where $*$ donates the conjugate, $\mathbf{v}_{k,i,j}$ is the k-th column of $\mathbf{V}_{i,j}$ and $V_{i,j}(k,k)$ represents the k-th diagonal element of $\mathbf{V}_{i,j}$. Applying EISS to each channel, we obtain the estimated $\hat{\mathbf{s}}_{i,j}$.

4. Experiments

In this section, we evaluated our proposed method's separation performance and sound quality in a double-talk environment. We compared the performance of SBSS-AEC with band-to-band filter and crossband filter using EISS. Regarding performance metrics, true echo return loss enhancement (tERLE) is used to measure separation performance. Besides, we use perceptual evaluation of speech quality (PESQ) [12] and short time objective intelligibility (STOI) [13] to measure the quality of the sound.

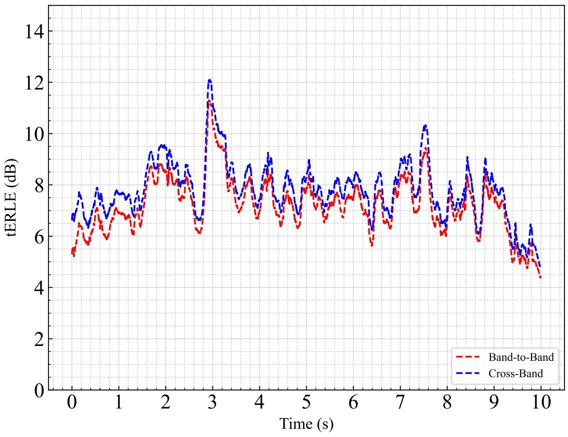

Figure 1: tERLE comparison of EISS methods with band-to-band and crossband filters

Table 1: Performance of EISS

T_{60}	Algorithm	PESQ	STOI	tERLE
300 ms	Band-to-Band	1.831	0.841	7.821
	Cross-Band ($P=1$)	1.880	0.853	8.444
	Cross-Band ($P=2$)	1.882	0.855	8.559
600 ms	Band-to-Band	1.390	0.565	5.326
	Cross-Band ($P=1$)	1.382	0.570	5.941
	Cross-Band ($P=2$)	1.413	0.571	6.064

4.1 Experimental Setup

In our simulation, 30 clean reading speech signals were randomly selected from the Deep Noise Suppression (DNS) challenge dataset [14]. All the signals have a length of 10 seconds, and the sampling rate is 16 kHz.

To simulate a realistic acoustic environment, this study generates room impulse response (RIR) based on the Image Source Method (ISM) [15]. The experimental setup specifies a room with dimensions of 8m × 8m × 3m and the reverberation time (T_{60}) of 300 ms and 600 ms. The microphone array is positioned at the center of the room at $(4.0, 4.0, 1.0)$, the loudspeaker is located at $(1.0, 4.0, 2.5)$, and the near-end speaker is positioned at $(4.77, 4.64, 1.51)$. The generated RIR contains 8192 samples and is normalized so that its maximum amplitude is limited to 0.6, ensuring the signal amplitude is suitable for subsequent simulation experiments.

For short-time analysis, the Hanning window is used and a frame length of 1024 points is adopted with a 75% overlap between adjacent frames. In the EISS process, the forgetting factor α is set to 0.992, and the shape parameter β is set to 0.4. We designed two experiments to evaluate our proposal. In the first experiment, the CTF filter length L is set to 5, and the number of crossband filters P is set as 2. In the second experiment, we set the number of cross-band filters to 2 and 4 respectively while changing the T_{60} from 300 ms to 600 ms. The CTF filer length L is increased to 7 when T_{60} is 600 ms.

4.2 Results

Figure 1 compares the tERLE of the conventional band-to-band filter-based AEC method with the proposed crossband filter-based EISS-AEC method under 300 ms reverberation. The signal-to-echo ratio (SER) of the signal is 0 dB. As plotted in the figure, our proposed method achieved an overall improvement in tERLE by introducing the cross-band filter into the separation.

In the second experiment, we test the performance of the proposed method in different reverberation times and P. The results are obtained as the average of three experiments, each comprising 30 data sets. The experiment results are shown in Table 1. The metrics comparison between the Band-to-Band filter and Cross-Band filter demonstrates that the precise mixing model of our proposal enhances the performance. However, the performance of the Cross-Band filter remains largely consistent across different P settings, suggesting that power leakage primarily occurs in adjacent subbands. To balance performance and latency, setting $P = 1$ will be an optimal choice according to this experiment.

5. Conclusions

In this paper, we incorporated a cross-band filter into the EISS algorithm to better model the echo in a reverberant environment. As the extended filter is very long, we used our previous EISS algorithm to update it. We conducted experiments in an AEC environment. The results confirmed that the cross-band filter performs better than the band-to-band filter in the EISS algorithm.

References

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

[2] 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.

[3] Y. Avargel and I. Cohen, “System identification in the short-time Fourier transform domain with crossband filtering,” *IEEE Trans. ASLP*, vol. 15, no. 4, pp. 1305–1319, 2007.

[4] T. Rosenbaum, I. Cohen, and E. Winebrand, “Crossband filtering for weighted prediction error-based speech dereverberation,” *Appl. Sci.*, vol. 13, no. 17, pp. 9537, 2023.

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

[6] 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. ICASSP*, 2024, pp. 756–760.

[7] 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 Trans. ASLP* 2024.

[8] L. Zhang, X. Wang, Y. Yang, T. Ueda, S. Makino, and J. Chen, ”Heavy-tailed Distributions-Based Online Semi-blind Source Separation for Nonlinear Echo Cancellation,” in *Proc. APSIPA ASC*, 2024, in press.

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

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

[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. ASLP*, vol. 19, no. 3, pp. 583–599, 2010.

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

[13] 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. ICASSP*, 2010, pp. 4214–4217.

[14] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, et al., “The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results,” in *Proc. ICASSP*, 2022.

[15] 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, 1979.