LOW ALGORITHMIC DELAY IMPLEMENTATION OF CONVOLUTIONAL BEAMFORMER FOR ONLINE JOINT SOURCE SEPARATION AND DEREVERBERATION

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

[1]Graduate School of Information, Production and Systems,
Waseda University, Kitakyushu, Japan
[2]Center of Intelligent Acoustics and Immersive Communications,
Northwestern Polytechnical University, Xi'an, China

ABSTRACT

Blind-audio-source-separation (BASS) techniques, particularly those with low latency, play an important role in a wide range of real-time systems, e.g., hearing aids, in-car hand-free voice communication, real-time human-machine interaction, etc. Most existing BASS algorithms are deduced to run on batch mode, and therefore large latency is unavoidable. Recently, some online algorithms were developed, which achieve separation on a frame-by-frame basis in the short-time-Fourier-transform (STFT) domain and the latency is significantly reduced as compared to those batch methods. However, the latency with these algorithms may still be too long for many real-time systems to bear. To further reduce latency while achieving good separation performance, we propose in this work to integrate a weighted prediction error (WPE) module into a non-causal sample-truncating-based independent vector analysis (NST-IVA). The resulting algorithm can maintain the algorithmic delay as NST-IVA if the delay with WPE is appropriately controlled while achieving significantly better performance, which is validated by simulations.

Index Terms— Independent vector analysis, weighted prediction error, non-causal sample truncating technique, algorithmic delay.

1. INTRODUCTION

Blind audio source separation (BASS) refers to the problem of separating audio source signals from observed mixtures with minimal prior information [1–4]. Many methods have been developed to tackle this problem, among which the so-called independent vector analysis (IVA) [5, 6] has been widely investigated and has demonstrated promising separation performance. Originally, IVA-based algorithms are deduced to run on batch mode, and therefore large latency is unavoidable, which is unacceptable in most real applications. To achieve low latency, online versions of IVA are developed [7–10]. This type of algorithms achieves separation on a frame-by-frame basis in the short-time-Fourier-transform (STFT) domain and the latency is significantly reduced as compared to those batch methods. However, the delay introduced by these algorithms may still be too long for many real-time applications since it depends on the frame length. The frame length in these algorithms has to be longer than the room impulse responses to achieve reasonably good separation performance.

To address this issue, a group of methods called convolutional beamformer (CBF) [12–14], which combine IVA and weighted prediction error (WPE) techniques [15, 16], is extended to the online version [17–19]. By integrating WPE, the CBF methods can process both current and past frames to mitigate the impact of reverberation even when the STFT frame length is short [17]. A drawback of this type of algorithms is that they are computationally expensive as the convolutional operation uses a number of frames to achieve dereverberation. In this work, we introduce a method that combines online CBF [17] and non-causal sample-truncating-based independent vector analysis (NST-IVA) [11]. This method uses a long STFT analysis window to reduce the length of convolutional operation in updating dereverberation filters while maintaining low algorithmic delay by truncating the non-causal samples of the process filters. Simulation results show that the proposed system is able to reduce the algorithmic delay to as low as 4 ms while producing better separation performance than its conventional counterparts.

This work was supported by JSPS KAKENHI Grant Number 23H03423.

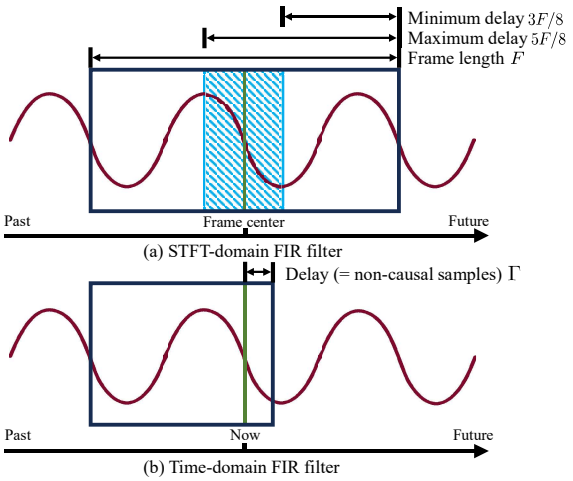

Fig. 1. Illustration of algorithmic delay of the STFT- (with window shift size of $F/4$) and time-domain methods.

2. NON-CAUSAL SAMPLE-TRUNCATING-BASED IVA

2.1. Algorithmic delay description

The algorithmic delay of the STFT- and time-domain algorithms is illustrated in Fig. 1. In the STFT-domain, due to the use of overlap-add technique, the output at time t is affected by multiple frames. To clarify the discussion, we calculate the algorithmic delay of an output point from the frame with the closest center to it. For example, as illustrated in Fig. 1(a), if we assume that the frame size is F and the overlap rate is 75%, the algorithmic delay of the samples within the light blue box depends on the frame noted by the black box. The average delay of these samples is F/2 since STFT cannot be performed until all samples in this frame are received. In other words, the algorithmic delay of an STFT-domain filter is related to the analysis window length. In comparison, for the time-domain algorithms, as illustrated in Fig. 1(b), the algorithmic delay of the filter depends

only on its non-causal sample number Γ. If truncating the non-causal components of a time-domain filter is feasible, its algorithmic delay can be reduced.

2.2. Conventional online IVA algorithms

Suppose that there are N sources in the sound field and we use an array of M microphones to pick up the signals. The observation signals in the STFT domain can be expressed as

$$\mathbf{x}(i,f) = \mathbf{A}(i;f)\mathbf{s}(i,f), \tag{1}$$

where

$$\mathbf{x}(i,f) = [X_1(i,f)\ \ X_2(i,f)\ \ \cdots\ \ X_M(i,f)]^{\mathsf{T}} \in \mathbb{C}^{M\times 1}, \tag{2}$$

$$\mathbf{s}(i,f) = [S_1(i,f)\ \ S_2(i,f)\ \ \cdots\ \ S_N(i,f)]^{\mathsf{T}} \in \mathbb{C}^{N\times 1}, \tag{3}$$

are the observed and source signal vectors respectively, f is the STFT bin index, i is the time frame index, $\mathbf{A}(i;f) \in \mathbb{C}^{M\times N}$ is the mixing matrix, and $(\cdot)^{\mathsf{T}}$ denotes the transpose operation. In this paper, we only consider the case with two microphones and two sources, i.e., $N = M = 2$. In our future work, we will investigate the causality of the separation filter with more input channels and apply this method to a larger microphone array. Assume that the $\mathbf{A}(i;f)$ matrix is of full rank, the source signals can be estimated with a linear filter, i.e.,

$$\mathbf{y}(i,f) = \mathbf{W}(i;f)\mathbf{x}(i,f), \tag{4}$$

where $\mathbf{W}(i;f) = \mathbf{A}^{-1}(i;f)$ is the separation matrix, $\mathbf{y}(i,f) = [Y_1(i,f)\ \ Y_2(i,f)\ \ \cdots\ \ Y_N(i,f)]^{\mathsf{T}} \in \mathbb{C}^{N\times 1}$ is an estimate of the source signals vector.

The algorithmic delay in the STFT domain is bounded to the STFT frame length. One way to reduce the delay is through truncation [11]. By inverse discrete Fourier transform (IDFT), the separation matrix $\mathbf{W}(i;f)$ is converted back to the time domain as

$$w_{n,m}(i;\tau) = \frac{1}{F}\sum_{f=-F/2}^{F/2-1} W_{n,m}(i;f)e^{j2\pi f\tau/F}, \tag{5}$$

where $W_{n,m}(i;f)$ is the (n,m)th element of $\mathbf{W}(i;f)$, and $w_{n,m}(i;\tau)$ corresponds to the time-domain FIR filter coefficient of nth source, mth input channel with the discrete-time index $\tau \in [-F/2, F/2-1]$. Then, the separation process is expressed as

$$y_n(t) = \sum_{\tau=-F/2}^{F/2-1}\sum_{m=1}^{M} w_{n,m}(i;\tau)x_m(t-\tau), \tag{6}$$

where $y_n(t)$ is the time-domain estimated signal of nth source at time t, $x_m(t)$ denotes the signal observed at the mth microphone. As proved in [11], the ideal separation filter $w_{n,m}(i;\tau)$ is a causal filter when $N = M = 2$. Hence, the algorithm delay can be reduced through truncation. Suppose that a total of Γ non-causal samples are truncated, the separated signal is expressed as

$$y_n(t) = \sum_{\tau=-F/2+\Gamma}^{F/2-1}\sum_{m=1}^{M} w_{n,m}(i;\tau)x_m(t-\tau). \tag{7}$$

The algorithm delay is then shortened from $F/2$ to $F/2-\Gamma$.

3. PROPOSED METHOD

3.1. System structure

The flowchart of the proposed algorithm is shown in Fig. 2. To further improve the BASS performance in heavy reverberation while maintaining a low algorithmic delay, a parallel processing structure similar to the one in NST-IVA [11] is used. This structure updates the joint separation and dereverberation filters by the conventional online CBF in the STFT-domain [17]. After updating, the filters are transformed back to the time-domain to separate new observed signals directly.

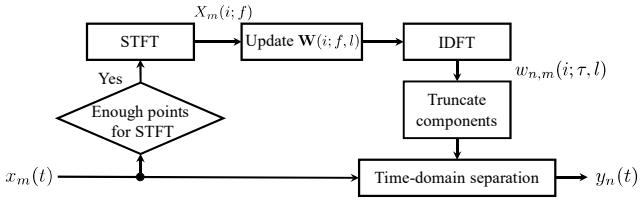

Fig. 2. Flowchart of the proposed algorithm.

3.2. Filters updating in the STFT domain

3.2.1. Problem formulation and probabilistic model

When the room impulse response is much longer than the STFT frame length, the instantaneous mixing model is not sufficient. Therefore, we adopt the convolutional transfer function (CTF) model in CBF [12, 13] to model the problem, i.e.,

$$\mathbf{x}(i,f) = \sum_{l_A=0}^{L_A-1} \mathbf{A}(i;f,l_A)\mathbf{s}(i-l_A,f), \tag{8}$$

where $\mathbf{A}(i;f,l_A) \in \mathbb{C}^{M\times N}$ is the convolutional mixing matrix at time lag l_A, and L_A is the order of the mixing filters. Correspondingly, the convolutional filters are used for simultaneous dereverberation and separation, i.e.,

$$\mathbf{y}(i,f) = \mathbf{W}(i;f,0)\mathbf{x}(i,f) + \sum_{l=D}^{L+D-1} \mathbf{W}(i;f,l)\mathbf{x}(i-l,f), \tag{9}$$

where $\mathbf{W}(i;f,l) \in \mathbb{C}^{N\times M}$ is a filter with time lag l, L is the total orders of filters, and D is a delay parameter introduced to prevent from distortion [15]. Based on the online source-wise factorization of CBF [17], the filters in (9) can be decomposed into the following two processes:

$$\mathbf{z}_n(i,f) = \mathbf{x}(i,f) - \mathbf{G}_n^{\mathsf{H}}(i;f)\overline{\mathbf{x}}(i,f), \tag{10}$$

$$Y_n(i,f) = \mathbf{q}_n^{\mathsf{H}}(i;f)\mathbf{z}_n(i,f). \tag{11}$$

The first process in (10) corresponds to the dereverberation process of the nth source, where $\mathbf{G}_n(i;f)$ and $\mathbf{z}_n(i,f)$ are the dereverberation filter and the output signal, $(\cdot)^{\mathsf{H}}$ represents conjugate transpose and $\overline{\mathbf{x}}(i,f) = [\mathbf{x}^{\mathsf{T}}(i-D,f)\ \ \mathbf{x}^{\mathsf{T}}(i-D-1,f)\ \ \cdots\ \ \mathbf{x}^{\mathsf{T}}(i-D-L+1,f)]^{\mathsf{T}}$ is the vector containing past samples of the mixture signals. The second process in (11) corresponds to the separation process, which uses filter $\mathbf{q}_n(i;f)$ to extract the nth source signal. Substituting (10) and (11) into (9), we obtain the relation between $\mathbf{W}(i;f,l)$ and $\mathbf{G}_n(i;f)$, $\mathbf{q}_n(i;f)$, i.e., $\mathbf{w}_n(i;f,0) = \mathbf{q}_n(i;f)$ and $[\mathbf{w}_n^{\mathsf{T}}(i;f,D),\cdots,\mathbf{w}_n^{\mathsf{T}}(i;f,L+D-1)]^{\mathsf{T}} = -\mathbf{G}_n(i;f)\mathbf{q}_n(i;f)$, where $\mathbf{w}_n(i;f,l)$ is the nth column of $\mathbf{W}^{\mathsf{H}}(i;f,l)$. Note that even though an STFT-domain separated signal component $Y_n(i,f)$ is produced in the updating process of $\mathbf{W}(i;f,l)$, this component is not be transformed to the time-domain as the final separated signal.

To deal with the joint dereverberation and source separation problem, the CBF method assumes that the source signal in the STFT-domain follows a Gaussian distribution with zero mean and time-dependent variance $\sigma_n(i) = \mathbb{E}(|S_n(i,f)|^2)$, i.e.,

$$S_n(i,f) \sim \mathcal{N}(0,\sigma_n(i)). \tag{12}$$

To estimate the adaptive filters to count for time-variant effects, the online version of CBF [17] adds a forgetting factor β to the

conventional CBF's negative log-likelihood function as

$$\mathcal{L}(\boldsymbol{\Theta}(i)) = -2\sum_f \log|\det \mathbf{Q}(i;f)| + \frac{\sum_{n,f,i'\le i}\beta^{i-i'}\left(\log\sigma_n(i')+\frac{|Y_n(i',f)|^2}{\sigma_n(i')}\right)}{\sum_{i'\le i}\beta^{i-i'}}, \quad (13)$$

where $\mathbf{Q}(i;f) = [\mathbf{q}_1(i;f)\ \ \mathbf{q}_2(i;f)\ \ \cdots\ \ \mathbf{q}_N(i;f)]^{\mathsf{H}} \in \mathbb{C}^{N\times M}$ is the separation matrix, and $\Theta(i) = \{\Theta_\sigma(i), \Theta_{\mathbf{G}}(i), \Theta_{\mathbf{Q}}(i)\}$ is the unknown parameter set, $\Theta_\sigma(i) = \{\sigma_n(i)\}$, $\Theta_{\mathbf{G}}(i) = \{\mathbf{G}_n(i;f)\}$, and $\Theta_{\mathbf{Q}}(i) = \{\mathbf{Q}_n(i;f)\}$. Each parameter set can be updated iteratively using the coordinate ascent method [20].

3.2.2. *Update of* $\Theta_\sigma(i)$

The process of filter update begins with calculating the estimated source signal $\mathbf{y}(i,f)$ using (10) and (11) with the filters $\mathbf{G}_n(i-1;f)$ and $\mathbf{Q}(i-1;f)$ being updated in the previous frame. Then, online CBF estimates the variance as

$$\sigma_n(i) \leftarrow \sum_{f=-F/2}^{F/2-1} |Y_n(i,f)|^2/F. \quad (14)$$

3.2.3. *Update of* $\Theta_{\mathbf{G}}(i)$

Following the method in [17], one can update $\Theta_{\mathbf{G}}(i)$ through minimizing (13) while fixing $\Theta_\sigma(i)$ and $\Theta_{\mathbf{Q}}(i)$, i.e,

$$\mathbf{G}_n(i;f) \leftarrow \mathbf{R}_n^{-1}(i;f)\mathbf{P}_n(i;f), \quad (15)$$

where

$$\mathbf{R}_n(i;f) = \beta\mathbf{R}_n(i-1;f) + \frac{\overline{\mathbf{x}}(i,f)\overline{\mathbf{x}}^{\mathsf{H}}(i,f)}{\sigma_n(i)}, \quad (16)$$

$$\mathbf{P}_n(i;f) = \beta\mathbf{P}_n(i-1;f) + \frac{\overline{\mathbf{x}}(i,f)\mathbf{x}^{\mathsf{H}}(i,f)}{\sigma_n(i)}, \quad (17)$$

are two spatio-temporal covariance matrices.

To achieve real-time processing, the matrix inversion lemma [21] is applied to promote the computational efficiency of (15). Hence, the calculation of $\mathbf{R}_{i,f,n}^{-1}$ can be written as

$$\mathbf{k}_n(i;f) \leftarrow \frac{\mathbf{R}_n^{-1}(i-1;f)\overline{\mathbf{x}}(i,f)}{\beta\sigma_n(i) + \overline{\mathbf{x}}^{\mathsf{H}}(i,f)\mathbf{R}_n^{-1}(i-1;f)\overline{\mathbf{x}}(i,f)}, \quad (18)$$

$$\mathbf{R}_n^{-1}(i;f) \leftarrow \frac{\mathbf{R}_n^{-1}(i-1;f) - \mathbf{k}_n(i;f)\overline{\mathbf{x}}^{\mathsf{H}}(i,f)\mathbf{R}_n^{-1}(i-1;f)}{\beta}, \quad (19)$$

where $\mathbf{k}_n(i;f)$ is the Kalman gain. Substituting (17) and (19) into (15) gives the following online update rule of $\mathbf{G}_n(i;f)$:

$$\mathbf{G}_n(i;f) \leftarrow \mathbf{G}_n(i-1;f) + \mathbf{k}_n(i;f)\mathbf{z}_n^{\mathsf{H}}(i,f). \quad (20)$$

3.2.4. *Update of* $\Theta_{\mathbf{Q}}(i)$

If fixing other parameters, the cost function in (13) degenerates to the one used in the online AuxIVA. Hence, the separation matrix $\mathbf{Q}_n(i;f)$ can be estimated through the iterative source steering (ISS)-based updating rules [22], which has been used in many IVA-based methods [9, 23, 24]. After initializing the separation matrix of the current frame $\mathbf{Q}_n(i;f) = \mathbf{Q}_n(i-1;f)$, this matrix can be updated with an auxiliary vector $\mathbf{v}_k(i;f)$ as

$$\mathbf{Q}(i;f) = \mathbf{Q}(i;f) - \mathbf{v}_k(i;f)\mathbf{q}_k^{\mathsf{H}}(i;f). \quad (21)$$

This update process is repeated for $k = 1,\ldots,N$. To continue the updating, $\mathbf{v}_k(i;f)$ needs to be calculated. By substituting (21) into (13) and fixing other parameters, the update rules of $\mathbf{v}_k(i;f)$ can be derived as

$$V_{n,k}(i;f) = \begin{cases} 1-(\mathbf{q}_n^{\mathsf{H}}(i;f)\mathbf{U}_n(i;f)\mathbf{q}_n(i;f))^{-1/2}, & \text{if } n=k, \\ \frac{\mathbf{q}_n^{\mathsf{H}}(i;f)\mathbf{U}_n(i;f)\mathbf{q}_k(i;f)}{\mathbf{q}_k^{\mathsf{H}}(i;f)\mathbf{U}_n(i;f)\mathbf{q}_k(i;f)}, & \text{else,} \end{cases} \quad (22)$$

where $V_{n,k}(i;f)$ is the nth element of $\mathbf{v}_k(i;f)$,

$$\mathbf{U}_n(i;f) = \alpha\mathbf{U}_n(i-1;f) + (1-\alpha)\frac{\mathbf{z}_n(i,f)\mathbf{z}_n^{\mathsf{H}}(i,f)}{\sigma_n(i)}, \quad (23)$$

is the weighted covariance matrix of the signal after dereverberation updated in an autoregressive manner [8] with a forgetting factor α.

3.3. Time-domain implementation

To reduce the algorithmic delay, we convert the original STFT-domain convolutional filters $\mathbf{W}(i;f,l)$ back to the time domain as in the conventional NST-IVA. Then, once a new signal sample is accessible, the output signal is generated immediately without waiting for enough samples for the next STFT frame. The transformation of $\mathbf{W}(i;f,l)$ by IDFT can be expressed as

$$w_{n,m}(i;\tau,l) = \frac{1}{F}\sum_{f=-F/2}^{F/2-1} W_{n,m}(i;f,l)e^{j2\pi f\tau/F}, \quad (24)$$

where $W_{n,m}(i;f,l)$ is the (n,m)th element of $\mathbf{W}(i;f,l)$ and $w_{n,m}(i;\tau,l)$ corresponds to the time-domain FIR filter parameter. Since STFT is a linear process, the STFT-domain joint separation and dereverberation process in (9) is equal to the following time-domain processing:

$$y_n(t) = \sum_{\tau=-F/2}^{F/2-1}\sum_{m=1}^{M}\Big[w_{n,m}(i;\tau,0)x_m(t-\tau) + \sum_{l=D}^{L+D-1} w_{n,m}(i;\tau,l)x_m(t-\tau-l\delta)\Big], \quad (25)$$

where δ is the window shift length of the STFT in updating filters. Moreover, we consider the window function $h(\tau)$ of the STFT with the discrete time index τ and rewrite the time domain processing (25) as

$$y_n(t) = \sum_{\tau=-F/2}^{F/2-1}\sum_{m=1}^{M}\Big[h(\tau)w_{n,m}(i;\tau,0)x_m(t-\tau) + \sum_{l=D}^{L+D-1} h(\tau)w_{n,m}(i;\tau,l)x_m(t-\tau-l\delta)\Big]. \quad (26)$$

To reduce the algorithmic delay, the non-causal samples of the filter $w_{n,m}(i;\tau,l)$ need to be truncated. If $D\times\delta \ge F/2$, all non-causal samples fall only in the separation filter $w_{n,m}(i;\tau,0)$; so, in this work, we consider only truncating $w_{n,m}(i;\tau,0)$. Besides, since the first tap of the convolutional transfer function $\mathbf{W}(i;f,0)$ is equal to the instant separation matrix in online IVA [8] as shown in subsection 3.2, the causality of $w_{n,m}(i;\tau,0)$ can be proved in a same way as [11]. Hence, the non-causal samples of $w_{n,m}(i;\tau,0)$ can be truncated without suffering heavy performance degradation. Suppose that a total of Γ non-causal samples are truncated, the proposed FIR filter process (26) can be written as

$$y_n(t) = \sum_{m=1}^{M}\Big[\sum_{\tau=-F/2+\Gamma}^{F/2-1} h(\tau)w_{n,m}(i;\tau,0)x_m(t-\tau) + \sum_{\tau=-F/2}^{F/2-1}\sum_{l=D}^{L+D-1} h(\tau)w_{n,m}(i;\tau,l)x_m(t-\tau-l\delta)\Big]. \quad (27)$$

Table 1. The process setting and the algorithmic delay of studied methods

Method	Window length	Truncated points	Algorithmic delay
TD-IVA (32 ms)	64 ms	0	32 ms
FD-CBF (4 ms)	8 ms	0	4 ms
TD-CBF (Proposed, 32 ms)	64 ms	0	32 ms
TD-CBF (Proposed, 4 ms)	64 ms	448	4 ms

With this step, the algorithmic delay is shortened to $F/2 - \Gamma$ samples as in NST-IVA.

4. EXPERIMENTAL EVALUATION

4.1. Experiment setup

The clean speech signals for simulations are taken from the ATR Japanese Speech Database [25]. Each mixed signal is generated with two source signals arbitrarily selected from two speakers from the database. If the source signal is less than 20 s, it is concatenated with another selected source signal and truncated then so the overall length is 20 s. The *pyroomacoustics* Python package [26] is used to simulate room impulse responses (RIRs) where the room boundary coefficients are controlled so that the room reverberation time, i.e., T_{60}, is 600 ms. We simulate a two-element microphone array with a spacing of 2 cm to observe the signals. The proposed method is validated under two different pairs of incidence angles in the two-speaker scenarios: (30°, 90°) and (30°, 150°). For each condition, 12 simulations with different mixed signals are performed and the average results are used for evaluation and comparison. The distance between the microphone array center and the sources is 2 m. All signals are sampled at 16 kHz.

Two conventional methods are used as the baseline systems: NST-IVA [11] (denoted as TD-IVA) and the STFT-domain online CBF [17] (denoted as FD-CBF). The proposed time-domain implementation of CBF is denoted as TD-CBF. Hann window is used as the analysis window in STFT and the overlap ratio is set to 75%. The frame length, the number of truncated samples, and the corresponding algorithmic delay of the studied methods are shown in Table 1. The values of D and L in CBF are set to 2 and 10 for the case with 64-ms frame length while 8 and 10 for FD-CBF. Although setting a longer WPE filter length, i.e., a larger value of L, can help achieve better performance when using shorter STFT windows, this would increase the computation cost. Therefore, we only used the same WPE filter length for comparison. The forgetting factor of IVA α and WPE β are set to 0.99 and 0.999 respectively. All simulations were conducted on a workstation powered by Intel Xeon E3-1505M.

The improvement of source-to-distortion ratio (ΔSDR), source-to-interferences ratio (ΔSIR), and sources-to-artifacts ratio (ΔSAR) [27] are used as the metrics to evaluate the separation performance.

4.2. Experiment result

The ΔSDR results as a function of time for all the studied methods are plotted in Fig. 3. As seen, all the studied methods converge at 10 s. For a fair comparison of performance, we compute the average ΔSDR, ΔSIR, and ΔSAR after all the methods converge, i.e., the results for the first 10-s are discarded. The results are shown in Fig. 4.

As seen in Fig. 3, FD-CBF with an 8 ms analysis window takes only 3 seconds to converge while the other methods with a 64 ms analysis window take 6 seconds to converge. This shows that the methods converge slower with longer STFT window length than

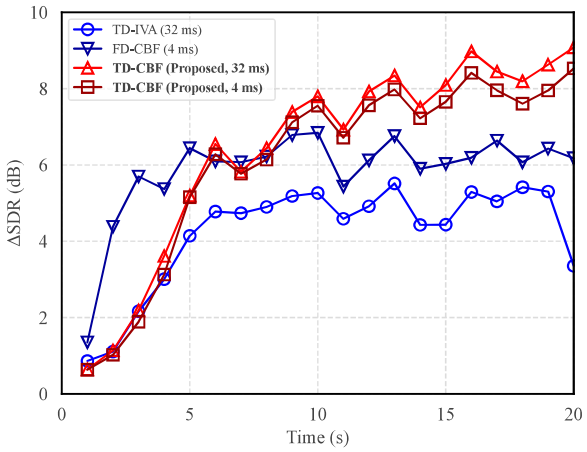

Fig. 3. Separation performance vs time.

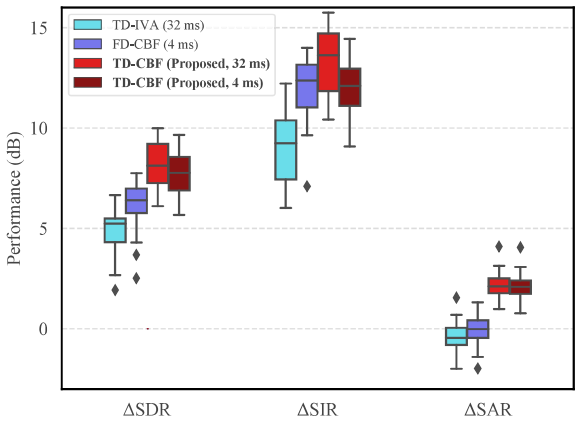

Fig. 4. Separation performance after convergence.

with shorter window length. According to this result, in practical systems, setting a relatively short window at the beginning and gradually increasing the length should be an appropriate way to achieve both good convergence speed and separation performance. From the results in Fig. 4, the proposed TD-CBF method (integrated with WPE dereverberation) yields much higher ΔSDR, ΔSIR, and ΔSAR results than TD-IVA after convergence. In comparison with FD-CBF, the proposed method can maintain the same algorithmic delay (4 ms) by truncating 448 non-causal samples, while using a relatively long STFT window length, which makes it more effective to deal with heavy reverberation. As a result, the proposed method also demonstrates better ΔSDR, ΔSAR than FD-CBF with the same algorithmic delay.

5. CONCLUSION

To achieve source recovery in reverberant environments with low latency, this paper developed an algorithm that combines the idea of non-causal sample truncation and WPE dereverberation method. Due to the application of non-causal sample truncation, the deduced algorithm is able to control the algorithmic delay as small as 4 ms. Meanwhile, due to the application of WPE, the algorithm is able to achieve better speech separation performance in strong reverberant environments in terms of source-to-distortion ratio and source-to-interferences ratio.

6. REFERENCES

[1] S. Makino, *Audio Source Separation*. Springer, 2018.

[2] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Wiley-IEEE Press., 2018.

[3] G. Huang, J. Benesty, and J. Chen, "Fundamental approaches to robust differential beamforming with high directivity factors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3074–3088, 2022.

[4] X. Wang, N. Pan, J. Benesty, and J. Chen, "On multiple input/binaural output antiphasic speaker signal extraction," in *Proc IEEE ICASSP*, 2023, pp. 1–5.

[5] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. ACSSC*, 2006, pp. 1393–1396.

[6] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.

[7] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. Circuits Syst. I*, vol. 57, no. 7, pp. 1431–1438, 2010.

[8] 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 IEEE HSCMA*, 2014, pp. 107–111.

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

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

[11] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *Proc IEEE ICASSP*, 2017, pp. 216–220.

[12] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2267–2282, Jul. 2020.

[13] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Computationally efficient and versatile framework for joint optimization of blind speech separation and dereverberation," in *Proc. Interspeech*, 2020, pp. 91–95.

[14] Y. Yang, X. Wang, A. Brendel, W. Zhang, W. Kellermann, and J. Chen, "Geometrically constrained source extraction and dereverberation based on joint optimization," in *Proc. EUSIPCO*, 2023, pp. 41-45.

[15] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[16] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[17] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online blind source separation based on joint optimization with blind dereverberation," in *Proc IEEE ICASSP*, 2021, pp. 506–510.

[18] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online source separation and noise reduction based on joint optimization with dereverberation," in *Proc. EUSIPCO*, 2021, pp. 1000–1004.

[19] T. Ueda and S. Makino, "Constant separating vector-based blind source extraction and dereverberation for a moving speaker," in *Proc. EUSIPCO*, 2023, pp. 930–934.

[20] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2010.

[21] P. S. Diniz *et al.*, *Adaptive filtering*. Springer, 1997.

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

[23] K. Goto, T. Ueda, L. Li, T. Yamada, and S. Makino, "Geometrically constrained independent vector analysis with auxiliary function approach and iterative source steering," in *Proc. EUSIPCO*, 2022, pp. 757–761.

[24] K. Mo, X. Wang, Y. Yang, T. Ueda, S. Makino, and J. Chen, "On joint dereverberation and source separation with geometrical constraints and iterative source steering," in *Proc. APSIPA ASC*, 2023, pp. 1138–1142.

[25] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[26] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," *Proc. IEEE ICASSP*, 2018, pp. 351–355.

[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.