LOW LATENCY ONLINE BLIND SOURCE SEPARATION BASED ON JOINT OPTIMIZATION WITH BLIND DEREVERBERATION

Tetsuya Ueda[1,2]*, Tomohiro Nakatani*[1]*, Rintaro Ikeshita*[1]*, Keisuke Kinoshita*[1]*, Shoko Araki*[1]*, and Shoji Makino*[2]

[1] NTT Corporation, Japan [2] University of Tsukuba, Japan

ABSTRACT

This paper presents a new low-latency online blind source separation (BSS) algorithm. Although algorithmic delay of a frequency domain online BSS can be reduced simply by shortening the short-time Fourier transform (STFT) frame length, it degrades the source separation performance in the presence of reverberation. This paper proposes a method to solve this problem by integrating BSS with Weighted Prediction Error (WPE) based dereverberation. Although a simple cascade of online BSS after online WPE upgrades the separation performance, the overall optimality is not guaranteed. Instead, this paper extends a recently proposed batch processing algorithm that can jointly optimize dereverberation and separation so that it can perform online processing with low computational cost and little processing delay ($<$ 12 ms). The results of a source separation experiment in a noisy car environment suggest that the proposed online method has better separation performance than the simple cascaded methods.

Index Terms— Blind source separation, blind dereverberation, online, independent vector analysis, real-time

1. INTRODUCTION

Speech enhancement is helpful for such applications as hearing aids and in-car communication systems (ICC), which transmit passengers' voices between the front and back seats for comfortable conversations in a vehicle [1]. However, such systems often need to eliminate highly nonstationary sounds, such as extraneous speaker voices, in real time with little delay. For achieving this, real-time blind source separation (BSS) is promising [2, 3]. BSS is a technique that separates individual source signals from microphone array inputs without any prior information about the source signals.

Several techniques for real-time (online) BSS have been investigated by modifying or extending non-real-time (batch or offline) BSS. The most commonly used approach for determined BSS (where the number of microphones equals that of the sources) is independent component analysis (ICA) [4], which achieves source separation by assuming the statistical independence between the sources. In frequency domain BSS, independent vector analysis (IVA) simultaneously solves separation and permutation alignment by assuming that the magnitudes of the frequency components originating from the same source tend to vary coherently over time [5]. Furthermore, auxiliary-function-based IVA (AuxIVA) [6, 7] and its online algorithm [8] have been proposed as a fast approach with rapid convergence and a low calculation cost.

In frequency domain BSS, we cannot neglect an algorithmic delay that is dependent on the short-time Fourier transform (STFT) frame length [9]. To achieve a sufficiently short delay for an ICC system, e.g., we need to use short STFT frames [10]. In contrast, such conventional online methods as AuxIVA assume that the STFT frame length must be longer than the reverberation time so that the source separation performance doesn't degrade. This trade-off can be solved by employing a dereverberation method, e.g., a Weighted Prediction Error (WPE) [11], and thus by removing the reverberation that continues longer than a frame. As a simple online approach using both separation and dereverberation, we can propose online-AuxIVA [8] cascaded after online-WPE [12, 13]. Although this method upgrades the separation performance, its overall optimality is not guaranteed because it separately optimizes WPE and BSS. For batch processing, although methods for jointly optimizing WPE and BSS have been reported [14, 15], they require the inversion of huge covariance matrices with a very special optimization structure of the WPE part, which requires high computational cost. Moreover, no online algorithm for it has been developed yet.

Recently, a new batch processing approach was proposed for joint optimization [16]. It factorizes the WPE part into sub-processings that separately dereverberate individual sources, and each sub-processing only requires the inversion of the covariance matrices with the same size and structure as those used for a conventional WPE. The total computational cost in the new joint method is much smaller than the conventional joint methods. Thus, this paper proposes a method to extend this new approach to online processing. Our proposal's advantages include its low computational complexity and high modularity, both of which allow us to combine conventional online algorithms for joint optimization with only minor modifications. We then evaluate the performance of the method by separation experiments on two-speaker mixtures, and show that it works effectively in real-time with low latency in a noisy car environment.

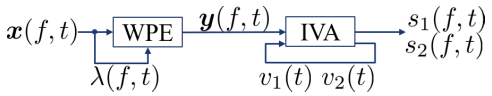
In the remainder of this paper, we describe the problem formulation in Section 2 and the baseline methods in Section 3. Our proposed technique is presented in Section 4. After a brief overview of related work in Section 5, experiments and conclusion are given in Sections 6 and 7.

2. PROBLEM FORMULATION

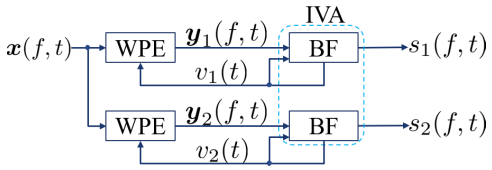
Suppose that N sources are captured by M microphones, and that the captured signals can be modeled at each time t and frequency f in the short-time Fourier transformation (STFT) domain:

$$\boldsymbol{x}(f,t) = \sum_{\tau=0}^{L_A-1} \boldsymbol{A}(f,\tau)\boldsymbol{s}(f,t-\tau) \tag{1}$$

where $\boldsymbol{s}(f,t) = [s_1(f,t),\ldots,s_N(f,t)]^{\mathsf{T}} \in \mathbb{C}^N$ and $\boldsymbol{x}(f,t) = [x_1(f,t),\ldots,x_M(f,t)]^{\mathsf{T}} \in \mathbb{C}^M$ are the vectors containing the



(a) Cascade optimization



(b) Joint optimization

Fig. 1: Cascade and joint optimization schemes, where $\lambda(f,t)$ and $v_n(t)$ were calculated from microphone signals, $\boldsymbol{x}(f,t)$ and IVA output, $s_n(f,t)$.

source and microphone signals, letting $(\cdot)^{\mathsf{T}}$ denote the transpose. $\boldsymbol{A}(f,\tau) \in \mathbb{C}^{M\times N}$ for $\tau = 0,\ldots,L_A-1$ is a convolutional transfer function matrix from the sources to the microphones, where L_A is the order of convolution. T and F are the total number of time frames and frequency bins. This paper assumes a determined case ($M = N$). Our aim here is to estimate a separation matrix that separates out individual source signals $\boldsymbol{s}(f,t)$ from $\boldsymbol{x}(f,t)$ in an online approach with a short STFT frame (= low algorithmic delay).

3. BASELINE METHODS

As a simple low-latency online approach, online-AuxIVA [8] that is cascaded after online-WPE [13] may be effective. Hereafter, we refer to this method as "WPE+IVA (cascade)". This paper takes online-AuxIVA and WPE+IVA (cascade) as baselines, and we briefly describe them in this section.

3.1. Online-AuxIVA

Online-AuxIVA is an effective algorithm to estimate the separation matrix [8]. It assumes that we can obtain source signals $\boldsymbol{s}(f,t)$ by multiplying a separation matrix to microphone signals $\boldsymbol{x}(f,t)$. The separation matrix can be obtained by Iterative Projection (IP) [7] and the recursive updates of the covariance matrices. A drawback of online-IVA is that when we shorten the STFT frame for low-latency processing, the source separation performance degrades in the presence of reverberation.

3.2. WPE+IVA (cascade)

As an effective algorithm to dereverberate microphone signals, several WPE techniques have been reported [11, 12, 13, 14]. WPE can perform online dereverberation by recursively updating a multichannel linear prediction filter to minimize the prediction error in each frequency bin.

Figure 1(a) illustrates the overall processing flow of WPE+IVA (cascade). By dereverberating signals prior to the application of IVA, WPE+IVA (cascade) may perform accurate separation even with a very short analysis window.

4. PROPOSED METHOD

One issue of WPE+IVA (cascade) is that its overall optimality is not guaranteed because the optimization is separately applied to WPE and BSS. In this section, we propose an online algorithm that jointly optimizes WPE and IVA.

4.1. Models of beamformers and source signals

To derive joint dereverberation and a source separation algorithm, we first assume that the relationship between $\boldsymbol{x}(f,t)$ and $\boldsymbol{s}(f,t)$ can be modeled by using a convolutional beamformer (CBF):

$$\boldsymbol{s}(f,t) = \boldsymbol{W}^{\mathsf{H}}(f,0)\boldsymbol{x}(f,t) + \sum_{\tau=D}^{D+L-1} \boldsymbol{W}^{\mathsf{H}}(f,\tau)\boldsymbol{x}(f,t-\tau), \quad (2)$$

where $\boldsymbol{W}(f,\tau)$ is a coefficient matrix, $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose, L is the length of the CBF, and D is the prediction delay. Under this assumption, $\boldsymbol{s}(f,t)$ can be estimated by identifying appropriate filter coefficients of a CBF.

Then we introduce a recently proposed technique[1], called source-wise factorization [16], to factorize a CBF into several WPE steps and an IVA step shown in Fig. 1(b). To explain this factorization, we first decompose the CBF in Eq. (2) into a set of CBFs, each of which independently estimates each source:

$$s_n(f,t) = \boldsymbol{w}_n^{\mathsf{H}}(f,0)\boldsymbol{x}(f,t) + \sum_{\tau=D}^{D+L-1} \boldsymbol{w}_n^{\mathsf{H}}(f,\tau)\boldsymbol{x}(f,t-\tau), \quad (3)$$

where $\boldsymbol{w}_n(f,\tau) \in \mathbb{C}^M$ for $\tau = 0, D, \ldots, D+L-1$ is the n-th column of $\boldsymbol{W}(f,\tau)$. Then, we factorize Eq. (3) into two sub-filters:

$$\boldsymbol{y}_n(f,t) = \boldsymbol{x}(f,t) - \boldsymbol{G}_n^{\mathsf{H}}(f)\overline{\boldsymbol{x}}(f,t), \quad (4)$$

$$s_n(f,t) = \boldsymbol{q}_n^{\mathsf{H}}(f)\boldsymbol{y}_n(f,t). \quad (5)$$

The first sub-filter, called a single-target WPE filter, dereverberates the n-th source with a prediction matrix $\boldsymbol{G}_n(f) \in \mathbb{C}^{ML\times M}$. $\overline{\boldsymbol{x}}(f,t) = [\boldsymbol{x}^{\mathsf{T}}(f,t-D),\ldots,\boldsymbol{x}^{\mathsf{T}}(f,t-D-L+1)]^{\mathsf{T}} \in \mathbb{C}^{ML}$ is a vector containing a past observation. The second sub-filter is a beamformer $\boldsymbol{q}_n(f) \in \mathbb{C}^M$ that extracts the n-th source signal. Note that we can easily show the equivalence between a CBF, Eq. (3), and the two forms of a CBF [16], namely, a pair of sub-filters, Eqs. (4) and (5).

Next, to derive an objective of the optimization, assume that $s_n(f,t)$ follows a zero-mean complex Gaussian distribution with variance $v_n(t) = \mathbb{E}[|s_n(f,t)|^2]$,

$$s_n(f,t) \sim \mathcal{N}_{\mathbb{C}}(0, v_n(t)). \quad (6)$$

Under this assumption, negative log-likelihood $\mathcal{I}$ given microphone signals $\mathcal{X} = \{x_m(f,t)\}_{m,f,t}$ becomes

$$\mathcal{I}(\mathcal{X}) \stackrel{c}{=} -2\sum_f \log\left|\det \boldsymbol{Q}(f)\right| + \frac{1}{T}\sum_{f,t,n}\left(\log v_n(t) + \frac{|s_n(f,t)|^2}{v_n(t)}\right), \quad (7)$$

where $\boldsymbol{Q}(f) = [\boldsymbol{q}_1(f),\ldots,\boldsymbol{q}_N(f)]$ and $\stackrel{c}{=}$ denotes equality up to constant terms.

[1] Conventionally researchers have used a different factorization technique for optimization [14, 15]. However, since this technique requires the calculation of a huge covariance matrix with a special structure, a step that complicates its online implementation, we do not adopt it in this paper.

4.2. Optimization by online approach

Now, we present our online algorithm for optimization. In the algorithm, we recursively update a set of parameters, $\theta_t = \{\mathcal{G}_t, \mathcal{Q}_t, \mathcal{V}_t\}$, where $\mathcal{G}_t = \{\boldsymbol{G}_n(f;t)\}_{f,n}$, $\mathcal{Q}_t = \{\boldsymbol{Q}(f;t)\}_f$, and $\mathcal{V}_t = \{v_n(t)\}_n$, in each frame. In addition, we modify the negative log likelihood function for online processing:

$$\mathcal{I}(\mathcal{X}_t) \stackrel{c}{=} -2\sum_f \log\left|\det \boldsymbol{Q}(f;t)\right| + \frac{1}{\sum_{t'\le t}\beta^{t-t'}} \sum_{f,t'\le t,n} \beta^{t-t'}\left(\log v_n(t') + \frac{|s_n(f,t')|^2}{v_n(t')}\right), \tag{8}$$

where $\mathcal{X}_t = \{\boldsymbol{x}_m(f,t')\}_{f,t'\le t,n}$ is the past and current microphone signals, and $0 < \beta < 1$ is a forgetting factor.

The negative log likelihood function (8) is decreased at each frame using a recursive coordinate descent method, which is comprised of the following three minimization steps:

$$\mathcal{V}_t \leftarrow \operatorname*{argmin}_{\mathcal{V}_t} \mathcal{I}(\mathcal{X}_t; \mathcal{G}_{t-1}, \mathcal{Q}_{t-1}, \mathcal{V}_t), \tag{9}$$

$$\mathcal{G}_t \leftarrow \operatorname*{argmin}_{\mathcal{G}_t} \mathcal{I}(\mathcal{X}_t; \mathcal{G}_t, \mathcal{Q}_{t-1}, \mathcal{V}_t), \tag{10}$$

$$\mathcal{Q}_t \leftarrow \operatorname*{argmin}_{\mathcal{Q}_t} \mathcal{I}(\mathcal{X}_t; \mathcal{G}_t, \mathcal{Q}_t, \mathcal{V}_t). \tag{11}$$

In the following, we describe the update equations for the above steps.

4.2.1. Update of $\mathcal{V}_t$

According to Eqs. (8) and (9), $\mathcal{V}_t$ can be updated by first estimating $\boldsymbol{y}_n(f,t)$ and $s_n(f,t)$ from $\boldsymbol{x}(f,t)$ based on Eqs. (4) and (5) using $\mathcal{G}_{t-1}$ and $\mathcal{Q}_{t-1}$ obtained in the previous time frame, and then calculating the variance of $s_n(f,t)$. Note that in the frequency domain ICA, the permutation of the separated components in each frequency is not uniquely determined. IVA solves this problem by averaging and dropping the frequency indices from $v_n(t)$:

$$v_n(t) \leftarrow \sum_{f=1}^{F} |s_n(f,t)|^2 / F. \tag{12}$$

4.2.2. Update of $\mathcal{G}_t$

By fixing $\mathcal{V}_t$, Eq. (8) can be minimized (without depending on $\mathcal{Q}_{t-1}$) by updating $\mathcal{G}_t$, as a previous work did [16]:

$$\boldsymbol{G}_n(f;t) = \boldsymbol{R}_n^{-1}(f;t)\boldsymbol{P}_n(f;t), \tag{13}$$

where $\boldsymbol{R}_n(f;t)$ and $\boldsymbol{P}_n(f;t)$ are spatio-temporal covariance matrices in the recursive form. They are derived from Eq. (8):

$$\boldsymbol{R}_n(f;t) = \beta\boldsymbol{R}_n(f;t-1) + \frac{\overline{\boldsymbol{x}}(f,t)\overline{\boldsymbol{x}}^{\mathsf{H}}(f,t)}{v_n(t)}, \tag{14}$$

$$\boldsymbol{P}_n(f;t) = \beta\boldsymbol{P}_n(f;t-1) + \frac{\overline{\boldsymbol{x}}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t)}{v_n(t)}. \tag{15}$$

Online update equations in $\boldsymbol{G}_n(f;t)$ and $\boldsymbol{R}_n^{-1}(f;t)$ can be obtained by applying the matrix inversion lemma [17]:

$$\boldsymbol{K}(f,t) \leftarrow \frac{\boldsymbol{R}_n^{-1}(f;t-1)\overline{\boldsymbol{x}}(f,t)}{\beta v_n(t) + \overline{\boldsymbol{x}}^{\mathsf{H}}(f,t)\boldsymbol{R}_n^{-1}(f;t-1)\overline{\boldsymbol{x}}(f,t)}, \tag{16}$$

$$\boldsymbol{R}_n^{-1}(f;t) \leftarrow \frac{\boldsymbol{R}_n^{-1}(f;t-1) - \boldsymbol{K}(f,t)\overline{\boldsymbol{x}}^{\mathsf{H}}(f,t)\boldsymbol{R}_n^{-1}(f;t-1)}{\beta}, \tag{17}$$

$$\boldsymbol{G}_n(f;t) \leftarrow \boldsymbol{G}_n(f;t-1) + \boldsymbol{K}(f,t)\boldsymbol{y}_n^{\mathsf{H}}(f,t), \tag{18}$$

where $\boldsymbol{K}(f,t)$ is the Kalman gain. The above update equation is identical to that of the conventional online-WPE optimization [13], except that the variance is obtained not by microphone signal, $\boldsymbol{x}(f,t)$, but by the beamformer output (Eq. (12)). In the above equation, since $\boldsymbol{R}_n(f;t)$ is much smaller than that used for another approach of WPE and IVA joint optimization [15], the computational cost for calculating $\boldsymbol{R}_n^{-1}(f;t)$ can be very small. This is particularly important for implementing an online algorithm.

4.2.3. Update of $\mathcal{Q}_t$

To update $\mathcal{Q}_t$, the log likelihood can be rewritten:

$$\mathcal{I}(\mathcal{Q}_t) \stackrel{c}{=} \sum_{f,n} \|\boldsymbol{q}_n(f;t)\|^2_{\boldsymbol{\Sigma}_n(f,t)} - 2\sum_f \log\left|\det \boldsymbol{Q}(f;t)\right|, \tag{19}$$

where $\boldsymbol{\Sigma}_n(f,t)$ is a covariance matrix used for the optimization, and $\|\boldsymbol{q}\|^2_\Sigma = \boldsymbol{q}^H\Sigma\boldsymbol{q}$. In this paper, we calculate $\boldsymbol{\Sigma}_n(f,t)$:

$$\boldsymbol{\Sigma}_n(f,t) \leftarrow \alpha\boldsymbol{\Sigma}_n(f,t-L_b) + (1-\alpha)\cdot\frac{1}{L_b}\sum_{\tau=t-L_b+1}^{t}\frac{\boldsymbol{y}_n(f,\tau)\boldsymbol{y}_n^{\mathsf{H}}(f,\tau)}{v_n(\tau)}, \tag{20}$$

where we slightly modify the configuration of the recursive update, following a previous study [8], by setting a different forgetting factor $0 < \alpha < 1$ and introducing a block-based covariance update with block length L_b. Because the above likelihood function is identical to that for conventional IVA except that the covariance matrix is calculated based on source dependent dereverberated signal $\boldsymbol{y}_n(f,t)$, methods for updating $\mathcal{Q}_t$ can be derived in almost the same way as those for the conventional techniques [6, 7, 8]. After initializing $\boldsymbol{Q}(f;t) = \boldsymbol{Q}(f;t-1)$ at each frame, AuxIVA updates $\boldsymbol{q}_n(f;t)$ using IP [7] in an online algorithm [8] in each t, f, n :

$$\boldsymbol{q}_n(f;t) \leftarrow (\boldsymbol{Q}^{\mathsf{H}}(f,t)\boldsymbol{\Sigma}_n(f,t))^{-1}\boldsymbol{e}_n, \tag{21}$$

$$\boldsymbol{q}_n(f;t) \leftarrow \frac{\boldsymbol{q}_n(f,t)}{\sqrt{\boldsymbol{q}_n^{\mathsf{H}}(f,t)\boldsymbol{\Sigma}_n(f,t)\boldsymbol{q}_n(f,t)}}, \tag{22}$$

where $\boldsymbol{e}_n$ denotes the n-th column of the $M \times M$ identity matrix, and $\boldsymbol{\Sigma}_n(f,t)$ is a covariance weighted by scalar $1/v_n(t)$.

In summary, the proposed algorithm in each t, f, and n is composed of the following three steps:

1. Update $v_n(t)$ using Eq. (12).
2. Update $\boldsymbol{G}_n(f;t)$ using Eqs. (16)–(18).
3. Update $\boldsymbol{q}_n(f;t)$ using Eqs. (20)–(22).

5. RELATED WORK

Another low-latency online-IVA was proposed [18] by implementing the separation matrix as FIR filters and the truncated part of their

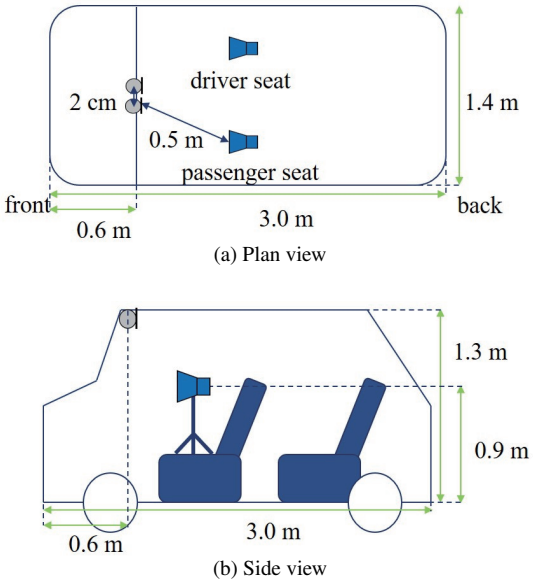


Fig. 2: Experimental sound source and microphone layout

non-causal components in the time domain. This method separates signals with an algorithmic delay less than 10 ms for hearing aids.

In contrast, our proposal differs in dereverberating microphone signals. WPE+IVA (joint) can estimate more dereverberated and cleaner source signals. No reports exist on jointing separation and dereverberation in a time domain. Comparing and jointing these methods are our future works.

6. EXPERIMENTAL EVALUATIONS

For our evaluation, we used set B of the ATR digital speech database [19]. It is composed of speech data from 10 speakers (six men and four women). By randomly selecting two different speakers from the database, we obtained 100 pairs of source signals. We generated microphone signals from the source signals by convoluting impulse responses recorded in a car and resampled the source and microphone signals at 16 kHz. We measured the impulse responses using a time-stretched pulse in a car. The recording environment is shown in Fig. 2. We set one speaker on the driver seat, another on a passenger seat, and microphones at a map lamp in the car. Reverberation time[2] T_{60} in the car was 48 ms. We compared three methods: an online-AuxIVA, WPE+IVA (cascade), and WPE+IVA (joint), which is our proposed method. We set the frame length and the shift at 8 ms and 4 ms based on the upper limit of the respective mouth-to-ear delays (= 12 ms) [10]. We took the average of the signal-to-distortion ratios (SDR), the signal-to-interference ratios (SIR), and the signal-to-noise ratios (SNR) as evaluation criteria [20]. We used a square root Hanning window for the STFT window by computational efficiency. We respectively set the other parameters $\{\alpha, \beta, L_b, D, L\}$ at $\{0.96, 0.9999, 2, 1, 5\}$. We initialized $\boldsymbol{Q}(f;0)$ and $\boldsymbol{R}_n^{-1}(f;0)$ as identity matrices, $\boldsymbol{G}_n(f;0)$ as a zero matrix, and $\Sigma_n(f,0)$ as $10^{-5} \times \boldsymbol{I}$, letting $\boldsymbol{I}$ be an identity matrix.

The first experiment was separation in a non-noisy environment. Fig. 3 compares the SIR improvements over time with each online method. The SIR improvement shows how much interference was

[2]This is much longer than an STFT frame in our experiments; thus we integrate dereverberation and source separation

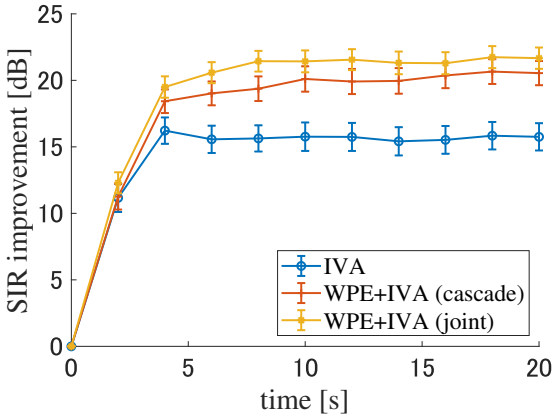


Fig. 3: Separation performance in comparison with each online method. Error bar denotes the $1.96\times$standard error in each time.

suppressed. We calculated the SIR every two-second period without overlap on signals to observe the variation in SIR over time. As the results show, the proposed method provided the highest SIR, although the difference between WPE+IVA (cascade) and WPE+IVA (joint) was not statistically significant.

The next experiment was separation in a noisy environment. The difference is that the microphone signals were generated by mixing the recorded diffuse noise and the mixed signal with input SNRs of 0 dB. The results are shown in Table 1. The proposed method outperformed all the baseline methods in terms of SDR, SIR, and SNR. Moreover, WPE+IVA (joint) showed significant SDR and SNR improvement, probably because it dereverberated microphone signals $\boldsymbol{x}(f,t)$ using the variance obtained not by $\boldsymbol{x}(f,t)$ but by beamformer output $\boldsymbol{s}(f,t)$. Targeting the source signal and removing the other noises may have caused the difference in SDR and SNR.

Table 1: Average SDR, SIR, and SNR improvements by each method. Bold font shows top scores. Scores in parentheses are $1.96\times$standard error.

Method	SDR [dB]	SIR [dB]	SNR [dB]
Online-AuxIVA	3.54 (0.59)	14.61 (0.98)	-1.68 (0.48)
WPE+IVA (cascade)	3.81 (0.59)	14.38 (0.91)	-1.44 (0.50)
WPE+IVA (joint)	**6.81 (0.50)**	**15.79 (0.66)**	**1.38 (0.47)**

Finally, we evaluated the computational efficiency in the proposed method. For 20 s of 2-ch input, the mean computation times of the proposed method were 2.7 s using python 3.7.7 on a computer with an Intel Core 3.6 GHz 1-core CPU. For 4-ch input, the times were 10.4 s, which corresponds to 2.08 ms ($=$ 10400 ms/5000 frames) for processing a frame on average. Thus, the total delay was 10.08 ms ($<$ 12 ms), including the algorithmic delay ($=$ 8 ms). This is sufficient performance for real-time separation.

7. CONCLUSIONS

We proposed low-latency online blind source separation. This method jointly optimized the dereverberation and separation online, with little algorithmic delay ($=$ 8 ms), and low computational cost ($=$ 10.08 ms). The experimental results confirmed that the proposed online method has better separation performance than the conventional online methods in non-noisy and noisy car environments.

8. REFERENCES

[1] R. Landgraf, J. Köhler-Kaeß, C. Lüke, O. Niebuhr, and G. Schmidt, "Can you hear me now? reducing the lombard effect in a driving car using an in-car communication system," in *Proc. Speech Prosody*, 2016, pp. 479–483.

[2] J. Antoni, "Blind separation of vibration components: Principles and demonstrations," *Mechanical Systems and Signal Processing*, vol. 19, no. 6, pp. 1166–1180, 2005.

[3] M. Zoulikha and M. Djendi, "A new regularized forward blind source separation algorithm for automatic speech quality enhancement," *Applied Acoustics*, vol. 112, pp. 192–200, 2016.

[4] A.Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[5] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.

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

[7] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-gaussian sources," in *Proc. LVA/ICA*. Springer, 2010, pp. 165–172.

[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. HSCMA*, 2014, pp. 107–111.

[9] D. Mauler and R. Martin, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," in *Proc. EUSIPICO*, 2007, pp. 222–226.

[10] A. Theiss, G. Schmidt, J. Withopf, and C. Lueke, "Instrumental evaluation of in-car communication systems," in *Speech Communication; 11. ITG Symposium*. VDE, 2014, pp. 1–4.

[11] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. ICASSP*, 2008, pp. 85–88.

[12] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. ICASSP*, 2009, pp. 3733–3736.

[13] J. Caroselli, I. Shafran, A. Narayanan, and R. Rose, "Adaptive multichannel dereverberation for automatic speech recognition.," in *Proc. Interspeech*, 2017, pp. 3877–3881.

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

[15] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *Proc. ICASSP*, 2018, pp. 31–35.

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

[17] S. Haykin, *Adaptive filter theory*, Pearson Education India, 2008.

[18] Masahiro Sunohara, Chiho Haruta, and Nobutaka 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. ICASSP*, 2017, pp. 216–220.

[19] 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.

[20] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.