NCSP'22
RISP International Workshop on Nonlinear circuits, Communications and Signal Processing 2022
February 28 - March 1

Blind Source Separation of Moving Sound Sources in Reverberant Indoor Environments

Ting Yu[1], Tetsuya Ueda[2] and Shoji Makino[1]

[1] Waseda University
2-7 Hibikino, Wakamatsu-ku, Kita-Kyushu, Fukuoka, Japan
E-mail: {yutingyu@toki, s.makino@}.waseda.jp

[2] University of Tsukuba
1–1–1 Tennodai, Tsukuba, Ibaraki, Japan
E-mail: t.ueda@mmlab.cs.tsukuba.ac.jp

Abstract

This paper explores whether the newly proposed online algorithm that jointly optimizes weighted prediction error (WPE) and independent vector analysis (IVA) works well in separating moving sound sources in reverberant indoor environments. The moving source is first fixed and then rotated 60 degrees in a room at a speed of less than 10 cm/s, while the other remains fixed. Through the comparison of the online-AuxIVA, online-WPE+IVA (separate) and online-WPE+IVA (joint) algorithms, we can conclude that the online-WPE+IVA (joint) method has the best separation performance when the sources are fixed, but online-WPE+IVA (separate) is more stable and has better performance when removing moving sources from the mixed sound.

1. Introduction

The separation of sound sources in reverberant indoor environments has important applications in scenarios such as video conferencing, speech enhancement and automatic speech recognition, and some solutions have been proposed using blind source separation (BSS) and joint optimization with blind dereverberation [1]. BSS is a technique that separates individual source signals from microphone array inputs without any prior information about the source signals. However, in real life, the observed signal is complex, sometimes the speaker of the target signal may move back and forth, at this time, it becomes more difficult to extract the target signal. Moreover, for real-time speech applications, we need to separate signals with little delay.

In frequency domain BSS, auxiliary-function-based independent vector analysis (AuxIVA) [2] and its online algorithm [3] have been proposed as a fast approach with rapid convergence and a low calculation cost. Although online-AuxIVA can realize moving source separation, it assumes that the short-time Fourier transform (STFT) frame length must be longer than the reverberation time so that the source separation performance does not degrade. However, frequency domain BSS such as IVA introduces an algorithmic delay that depends on the STFT frame length [4]. Therefore, to remove the reverberation that continues longer than a frame, employing a dereverberation method [5], e.g., a weighted prediction error (WPE) [6] has been proposed. To achieve low latency processing, cascading online-WPE based dereverberation [8] with an online-AuxIVA [3] has been effective [7]. But it does not guarantee overall optimality because the optimization is separately applied to WPE and BSS. Then, an online algorithm that jointly optimizes WPE and BSS was proposed to solve this problem [7]. In this method, it decomposes one dereverberation filter into several filters. With this technique, we can maintain separation performance with a short STFT frame (= low latency) and optimize WPE and IVA jointly in online with a low computational cost.

We noticed that [7] focused on the only in-car environment, where speakers are fixed and reverberation time is very short. Since this method [7] is based on online-AuxIVA, it is theoretically capable of separating moving sound sources with little latency as well. Therefore, we applied the recently proposed method [7] to moving source separation and evaluated the separation performance.

In the remainder of this paper, we describe the methods in Section 2. Experiments and conclusions are given in Sections 3 and 4.

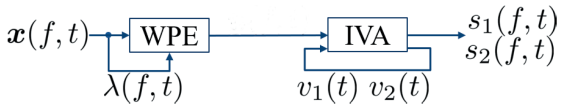
2. Method

In this paper, we assume a convolutive mixture model and a determined situation. Suppose N source signals are observed by M microphones ($N = M$), where $\boldsymbol{s}(f,t) = [s_1(f,t),\ldots,s_N(f,t)] \in \mathbb{C}^N$ and $\boldsymbol{x}(f,t) = [x_1(f,t),\ldots,x_M(f,t)] \in \mathbb{C}^M$ are the vectors containing the source and microphone signals. The observed signals can be modeled at each time t and frequency f in the STFT domain. We consider a relationship between n-th source signal $s_n(f,t)$ and observed signal $\boldsymbol{x}(f,t)$ as

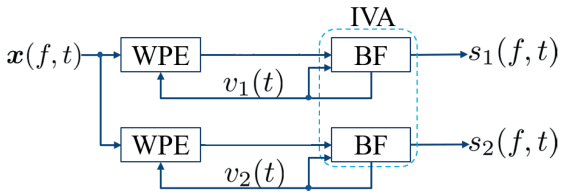
$$\boldsymbol{y}_n(f,t) = \boldsymbol{x}(f,t) - \boldsymbol{G}_n^{\mathsf{H}}(f)\overline{\boldsymbol{x}}(f,t), \tag{1}$$

$$s_n(f,t) = \boldsymbol{q}_n^{\mathsf{H}}(f)\boldsymbol{y}_n(f,t). \tag{2}$$

where $\boldsymbol{G}_n(f) \in \mathbb{C}^{ML\times M}$ is a prediction matrix which can dereverberate the n-th source and $\overline{\boldsymbol{x}}(f,t) = [\boldsymbol{x}^{\mathsf{T}}(f,t-D),\ldots,\boldsymbol{x}^{\mathsf{T}}(f,t-D-L+1)]^{\mathsf{T}} \in \mathbb{C}^{ML}$ is a vector containing a past observation. L is the length of the observation for using dereverberation, and D is the prediction delay. The separation filter $\boldsymbol{q}_n(f)$ is a beamformer that extracts the n-th source signal. $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose and $(\cdot)^{\mathsf{T}}$ denotes the transpose. For this separation model, we show three online algorithms.



(a) Cascade optimization schemes



(b) Joint optimization schemes

Figure 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, $s_n(f,t)$

2.1 Online-AuxIVA

In online-AuxIVA, the sources are estimated by a linear demixing process. This method [3] has a low computational cost but it will introduce a delay caused by STFT analysis which degrades the separation performance. On the other hand, if we shorten the STFT frame, source separation performance degrades by reverberation. In the above model, online-AuxIVA updates only $\boldsymbol{q}_n(f)$.

2.2 Online-WPE+IVA (separate)

It has been reported [6,8–10] that WPE is an effective algorithm that can de-reverberate microphone signals. As a simple approach which cascading online-AuxIVA and online-WPE [8] may maintain separation performance with shorting STFT frame. Figure 1(a) illustrates the overall processing flow of online-WPE+IVA (separate). In the above model, the online-WPE+IVA (separate) update $\boldsymbol{q}_n(f)$ and $\boldsymbol{G}_n(f)$ using [3] and [5] respectively. Note that online-WPE+IVA (separate) uses only one dereverberation filter $\boldsymbol{G}(f)$ instead of multiple filters $\boldsymbol{G}_n(f)$.

2.3 Online-WPE+IVA (joint)

The disadvantage of online-WPE+IVA (separate) is that its overall optimality is not guaranteed because the optimization is separately applied to WPE and BSS. Online-WPE+IVA (joint) [7] solves this problem by jointly optimizing WPE and BSS as shown in Fig. 1(b).

To derive an objective of the optimization, we assume that sound sources are mutual independent and $s_n(f,t)$ follows a zero-mean complex Gaussian distribution with variance $v_n(t) = E[|s_n(f,t)|^2]$. Under the above assumptions, given past and current microphone signals $\mathcal{X} = \{x_m(f,t)\}_{m,f,t}$ with forgetting factor $0 < \beta < 1$, negative log-likelihood $\mathcal{I}$ becomes:

$$\mathcal{I}(\mathcal{X}_t) \stackrel{c}{=} -2\sum_f \log\left|\det \boldsymbol{Q}(f;t)\right| + \frac{1}{\sum_{t'\le t}\beta^{t-t'}}\sum_{f,t'\le t,n}\beta^{t-t'}\left(\log v_n(t') + \frac{|s_n(f,t')|^2}{v_n(t')}\right), \tag{3}$$

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

We can decrease this function using a recursive coordinate descent method in each time t, that is, 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. It can be comprised of the following three minimization steps:

$$\mathcal{V}_t \leftarrow \operatorname*{argmin}_{\mathcal{V}_t} \mathcal{I}(\mathcal{X}_t; \mathcal{G}_{t-1}, \mathcal{Q}_{t-1}, \mathcal{V}_t), \tag{4}$$

$$\mathcal{G}_t \leftarrow \operatorname*{argmin}_{\mathcal{G}_t} \mathcal{I}(\mathcal{X}_t; \mathcal{G}_t, \mathcal{Q}_{t-1}, \mathcal{V}_t), \tag{5}$$

$$\mathcal{Q}_t \leftarrow \operatorname*{argmin}_{\mathcal{Q}_t} \mathcal{I}(\mathcal{X}_t; \mathcal{G}_t, \mathcal{Q}_t, \mathcal{V}_t). \tag{6}$$

According to Eqs. (3) and (4), $\mathcal{V}_t$ can be updated by first estimating $\boldsymbol{y}_n(f,t)$ and $s_n(f,t)$ from $\boldsymbol{x}(f,t)$ based on Eqs. (1) and (2) using $\mathcal{G}_{t-1}$ and $\mathcal{Q}_{t-1}$ obtained in the previous time frame, and then calculating the variance of $s_n(f,t)$. Moreover, IVA solves the problem that the permutation of the separated components in each frequency are not uniquely determined by averaging and dropping the frequency indices from $v_n(t)$:

$$v_n(t) \leftarrow \sum_{f=1}^{F} |s_n(f,t)|^2/F. \tag{7}$$

By fixing $\mathcal{V}_t$, Eq. (3) can be minimized (without depending on $\mathcal{Q}_{t-1}$) by updating $\mathcal{G}_t$, as a previous work did [10]:

$$\boldsymbol{G}_n(f;t) = \boldsymbol{R}_n^{-1}(f;t)\boldsymbol{P}_n(f;t), \tag{8}$$

where $\boldsymbol{R}_n(f;t)$ and $\boldsymbol{P}_n(f;t)$ are spatio-temporal covariance matrices in the recursive form which can be derived from Eq. (3):

$$\boldsymbol{R}_n(f;t) = \beta\boldsymbol{R}_n(f;t-1) + \frac{\overline{\boldsymbol{x}}(f,t)\overline{\boldsymbol{x}}^{\mathsf{H}}(f,t)}{v_n(t)}, \tag{9}$$

$$\boldsymbol{P}_n(f;t) = \beta\boldsymbol{P}_n(f;t-1) + \frac{\overline{\boldsymbol{x}}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t)}{v_n(t)}. \tag{10}$$

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 [11]:

$$\boldsymbol{K}(f,t) \leftarrow \frac{\boldsymbol{R}_n^{-1}(f;t-1)\overline{\boldsymbol{x}}(f,t)}{\beta v_n(t) + \overline{\boldsymbol{x}}^{\mathsf{H}}(f,t)\boldsymbol{R}_n^{-1}(f;t-1)\overline{\boldsymbol{x}}(f,t)}, \tag{11}$$

$$\boldsymbol{R}_n^{-1}(f;t) \leftarrow \frac{\boldsymbol{R}_n^{-1}(f;t-1) - \boldsymbol{K}(f,t)\overline{\boldsymbol{x}}^{\mathsf{H}}(f,t)\boldsymbol{R}_n^{-1}(f;t-1)}{\beta}, \tag{12}$$

$$\boldsymbol{G}_n(f;t) \leftarrow \boldsymbol{G}_n(f;t-1) + \boldsymbol{K}(f,t)\boldsymbol{y}_n^{\mathsf{H}}(f,t), \tag{13}$$

where $\boldsymbol{K}(f,t)$ is the Kalman gain.

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

$$\mathcal{I}(\mathcal{Q}_t) \stackrel{c}{=} \sum_{f,n} \|\boldsymbol{q}_n(f;t)\|^2_{\boldsymbol{\Sigma}_n(f,t)} - 2\sum_{f} \log\left|\det \boldsymbol{Q}(f;t)\right|, \tag{14}$$

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

$$\begin{aligned}\boldsymbol{\Sigma}_n(f,t) \leftarrow{}& \alpha \boldsymbol{\Sigma}_n(f,t-L_b) \\ &+ (1-\alpha)\cdot \frac{1}{L_b} \sum_{\tau=t-L_b+1}^{t} \frac{\boldsymbol{y}_n(f,\tau)\boldsymbol{y}_n^{\mathsf{H}}(f,\tau)}{v_n(\tau)},\end{aligned} \tag{15}$$

where the configuration of the recursive update is slightly modified following a previous study [3], by setting a different forgetting factor $0 < \alpha < 1$ and introducing a block-based covariance update with block length L_b.

After initializing $\boldsymbol{Q}(f;t) = \boldsymbol{Q}(f;t-1)$ at each frame, AuxIVA updates $\boldsymbol{q}_n(f;t)$ using Iterative Projection(IP) [2] in an online algorithm [3] in each t, f, n :

$$\boldsymbol{q}_n(f;t) \leftarrow (\boldsymbol{Q}^{\mathsf{H}}(f,t)\boldsymbol{\Sigma}_n(f,t))^{-1}\boldsymbol{e}_n, \tag{16}$$

$$\boldsymbol{q}_n(f;t) \leftarrow \frac{\boldsymbol{q}_n(f,t)}{\sqrt{\boldsymbol{q}_n^{\mathsf{H}}(f,t)\boldsymbol{\Sigma}_n(f,t)\boldsymbol{q}_n(f,t)}}, \tag{17}$$

where $\boldsymbol{e}_n$ denotes the n-th column of the $M \times M$ identity matrix.

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

1. Update $v_n(t)$ using (7).
2. Update $\boldsymbol{G}_n(f;t)$ using (11)-(13).
3. Update $\boldsymbol{q}_n(f;t)$ using (15)-(17).

3. Experimental evaluations

In order to evaluate the effectiveness of these three methods for separating moving sound sources, we conducted a source separation experiment. In this experiment, we obtained 10 pairs of source signals by randomly selecting two different speakers from the set B of the ATR digital speech database [12]. The distance spacing of microphones was set to 10 cm. We let the position of one of the sources remain fixed for the entire 20 seconds, while the other source is fixed in the first 10 seconds and then moves in the second 10 seconds, as shown in Fig. 2, where $src1$ represents the speaker whose position is fixed, and $src2$ represents the speaker whose position will change. In making sources, we used image method [13]. We generated reverberant signals of the fixed speaker using " RIR generator " [14]. For the moving sound sources, we generated reverberant signals using "signal generator" [15]. The experimental conditions are shown in Table 1.

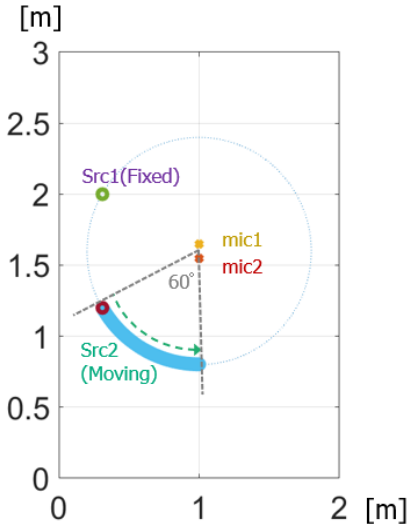


Figure 2: Layout of experimental environment

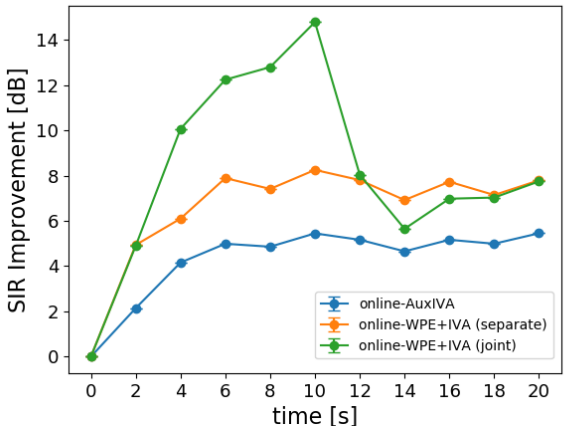
Table 1: Experimental conditions

STFT analysis window	Square root hanning
STFT frame, shift	32 ms, 16 ms
Dereverberation length(L)	5
Prediction delay(D)	1
Block size for IVA (L_b)	2
Forgetting factor for IVA (α)	0.96
Forgetting factor for WPE (β)	0.99
Initial value of $\boldsymbol{G}_n(f;0)$	Zero matrix
Initial value of $\boldsymbol{\Sigma}_n(f;0)$	Identity matrix
Initial value of $\boldsymbol{Q}(f;0)$	Identity matrix

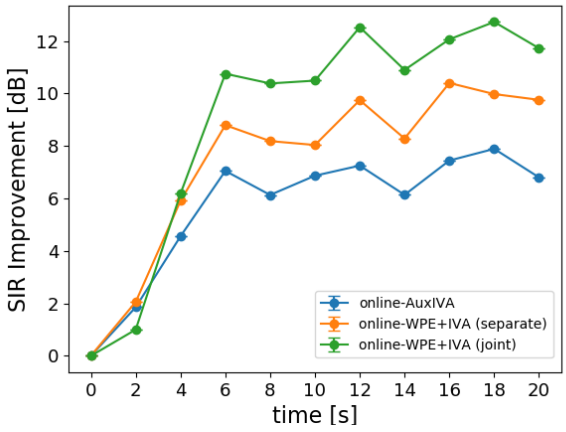
Figure 3 compares each online method's source-to-interference ratio (SIR) [16] improvements over two seconds. Figure 3 (a) shows the SIR improvements of $src1$ and Fig. 3(b) shows the SIR improvements of $src2$. As shown in Fig. 3(a), when $src2$ starts to move after 10 s, all three curves start to fall, with the curve of online-WPE+IVA (joint) dropping faster than the others. It shows that online-WPE+IVA (joint) cannot remove $src2$ well, causing the curve to drop very fast. Since $src1$ is permanently fixed, it is better to remove $src1$ from the mixed sound, so there is no dramatic drop in the three curves in Fig. 3(b). Moreover, it can be seen that, in general, online-WPE+IVA (joint) has the best separation performance when separating the mixed sound of two fixed-position sources in reverberant indoor environments.

4. Conclusions

In this paper, we evaluated the effectiveness of online-AuxIVA, online-WPE+IVA (separate) and online-WPE+IVA (joint) for separating moving sound sources. We conducted a separation experiment. The results show that online-WPE+IVA (joint) has the best separation performance when



(a) SIR improvements of $src1$ (Fixed)



(b) SIR improvements of $src2$ (Moving)

Figure 3: Separation performance in comparison with each online method

separating the mixed sound of two fixed-position sources in reverberant indoor environments. However, when separating the moving sound sources, online-WPE+IVA (joint) has poorer separation performance than online-WPE+IVA (separate).

References

[1] 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.

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

[3] 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.

[4] 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.

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

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

[7] 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. ICASSP*, 2021, pp. 506-510.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] L. Li, K. Koishida, S. Makino, "Online Directional Speech Enhancement Using Geometrically Constrained Independent Vector Analysis". in *Proc. Interspeech*, 2020, pp. 61-65.

[14] https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

[15] https://www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator

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