Computationally Efficient and Versatile Framework for Joint Optimization of Blind Speech Separation and Dereverberation

Tomohiro Nakatani, Rintaro Ikeshita, Keisuke Kinoshita, Hiroshi Sawada, Shoko Araki

NTT Corporation, Japan

Abstract

This paper proposes new blind signal processing techniques for optimizing a multi-input multi-output (MIMO) convolutional beamformer (CBF) in a computationally efficient way to simultaneously perform dereverberation and source separation. For effective CBF optimization, a conventional technique factorizes it into a multiple-target weighted prediction error (WPE) based dereverberation filter and a separation matrix. However, this technique requires the calculation of a huge spatio-temporal covariance matrix that reflects the statistics of all the sources, which makes the computational cost very high. For computationally efficient optimization, this paper introduces two techniques: one that decomposes the huge covariance matrix into ones for individual sources, and another that decomposes the CBF into sub-filters for estimating individual sources. Both techniques effectively and substantively reduce the size of the covariance matrices that must calculated, and allow us to greatly reduce the computational cost without loss of optimality.

Index Terms: Blind source separation, dereverberation, automatic speech recognition

1. Introduction

When a speech signal is captured by distant microphones, e.g., in a conference room, it often contains reverberation, diffuse noise, and voices of extraneous speakers. These components are detrimental to the intelligibility of the captured speech and often cause serious degradation in many speech applications, such as Automatic Speech Recognition (ASR).

Blind signal processing minimizes the aforementioned detrimental effects in the acquired signals without prior knowledge of the sources or the room acoustics. For reduction of extraneous speakers' voices, a number of techniques have been developed for blind source separation (BSS), including independent component analysis [1, 2], independent vector analysis [3, 4, 5], and spatial clustering-based time-frequency masking and beamforming [6, 7, 8]. It has also been empirically confirmed that BSS can perform denoising [9, 10]. For blind dereverberation (BDR), a Weighted Prediction Error minimization (WPE)-based technique [11, 12, 13] has been actively studied as an effective approach.

Techniques for jointly optimizing BSS and BDR based on a multi-input multi-output (MIMO) convolutional beamformer (CBF) have also been investigated [14, 15, 16, 17, 18, 19, 20, 21, 22]. For example, with certain techniques [16, 18, 19], a CBF is factorized into a multiple-target weighted prediction error (WPE) dereverberation filter and a separation matrix and jointly optimized. This approach, however, requires the calculation of a huge spatio-temporal covariance matrix that reflects the statistics of all the sources. This makes the computational cost very high and has inhibited wider use of these techniques.

To achieve computationally efficient optimization, we reintroduce two techniques that we have recently proposed to optimize a (non-blind) mask-based CBF [23, 24]. One is source-wise covariance decomposition, which decomposes a huge covariance matrix into smaller ones that correspond to individual sources, and the other is source-wise CBF factorization, which factorizes a CBF into a set of single-target WPE filters that correspond to individual sources and a separation matrix. We show that each technique can accomplish computationally much more efficient joint optimization without loss of optimality based on spatio-temporal covariance matrices that are calculated separately for individual sources.

In the remainder of this paper, after a brief overview on related work in section 2, we describe the problem formulation and a conventional joint optimization technique in sections 3 and 4. Our proposed techniques are presented in section 5. Finally, experiments and concluding remarks are given in sections 6 and 7.

2. Related work

The optimization techniques used in this paper were first proposed for (non-blind) optimization of a CBF under a condition where the time-frequency masks of the target signals are given or can be estimated [24]. We newly apply these techniques to blind signal processing and experimentally examine their effectiveness in this paper.

To avoid the calculation of a huge covariance matrix for the efficient optimization of a CBF, a different scheme has also been proposed [21, 25]. With it, a CBF is optimized without being factorized into WPE filters and a separation matrix. In contrast, the techniques for factorizing a CBF proposed in our paper enable more flexible control of the optimization. For example, computational efficiency can be further enhanced by adopting different iteration schemes to update BSS and BDR, as will be discussed in our experiments.

Note that the joint optimization problems solved in this paper and previous articles [16, 19, 21] are equivalent because they are based on the same CBF and optimization criteria in the same family. A major difference is how the CBF is parameterized. This results in different optimization algorithms.

3. 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:

$$\mathbf{x}_{t,f} = \sum_{\tau=0}^{L_A-1} \mathbf{A}_{\tau,f}\mathbf{s}_{t-\tau,f}, \tag{1}$$

where $\mathbf{s}_{t,f} = [s^{(1)}_{t,f}, \ldots, s^{(N)}_{t,f}]^\top \in \mathbb{C}^{N\times 1}$ and $\mathbf{x}_{t,f} = [x_{1,t,f}, \ldots, x_{M,t,f}]^\top \in \mathbb{C}^{M\times 1}$ are the vectors containing source and microphone signals, where $(\cdot)^\top$ denotes a non-conjugate transpose, $\mathbf{A}_{\tau,f} \in \mathbf{C}^{M\times N}$ for $\tau = 0, \ldots, L_A - 1$

is a convolutional transfer function matrix from the sources to the microphones. This paper assumes $M = N$, which is a determined case. To perform simultaneous dereverberation and source separation, we employ a CBF:

$$\mathbf{y}_{t,f} = \sum_{\tau=0}^{L-1} \mathbf{W}_{\tau,f}^{\mathsf{H}} \mathbf{x}_{t-\tau,f}. \tag{2}$$

where $(\cdot)^H$ denotes a conjugate transpose, and $\mathbf{W}_{\tau,f} \in \mathbb{C}^{M\times N}$ for $\tau = 0, \ldots, L-1$ is a coefficient matrix of the CBF. Next we model each source $s_{t,f}^{(i)}$ by a zero-mean complex Gaussian distribution with a time-frequency dependent variance $\lambda_{t,f}^{(i)}$. In addition, we assume that $s_{t,f}^{(n)}$ and $s_{t',f'}^{(n')}$ are mutually independent for $(t, f, n) \neq (t', f', n')$, and that there is a certain set of coefficients $\Theta_{\mathbf{W}} = \{\mathbf{W}_{\tau,f}\}$ that makes $\mathbf{y}_{t,f} = \mathbf{s}_{t,f}$. Then the log likelihood function can be written:

$$\mathcal{L}(\theta) = -\sum_{t,f,n} \left(\log \lambda_{t,f}^{(n)} + \frac{|y_{t,f}^{(n)}|^2}{\lambda_{t,f}^{(n)}} \right) + 2T \sum_{f} \log |\det \mathbf{W}_{0,f}|, \tag{3}$$

where T is the total number of time frames, $\theta = \{\Theta_{\mathbf{W}}, \Theta_\lambda\}$, and $\Theta_\lambda = \{\lambda_{t,f}^{(n)}\}$. A CBF can be optimized by estimating θ that maximizes the above log likelihood function.

Note that some of the assumptions introduced in the above formulation are not accurate. For example, the existence of the CBF that exactly recovers the sources is guaranteed only with over-determined cases ($M > N$) [26]. This means that the likelihood function is logical only in an approximate sense. In addition, $s_{t,f}^{(n)}$ inherently has temporal correlation within each short time duration. A technique is often introduced to prevent the inherent correlation from being decorrelated by the CBF, where we set the dereverberation goal to reduce only the late reverberation without changing the direct signal and the early reflections. This is simply done by introducing prediction delay D [11, 27] into the CBF:

$$\mathbf{y}_{t,f} = \mathbf{W}_{0,f}^{\mathsf{H}} \mathbf{x}_{t,f} + \sum_{\tau=D}^{L-1} \mathbf{W}_{\tau,f}^{\mathsf{H}} \mathbf{x}_{t-\tau,f}. \tag{4}$$

4. Conventional optimization method

Conventionally, researchers have derived techniques for optimizing a CBF based on a way of factorizing it [16, 19], referred to in this paper as source-packed factorization [24].

4.1. Source-packed CBF factorization

With this factorization, a CBF is factorized into two sub-filters:

$$\mathbf{z}_{t,f} = \mathbf{x}_{t,f} - \mathbf{G}_f^{\mathsf{H}} \overline{\mathbf{x}}_{t,f}, \tag{5}$$

$$\mathbf{y}_{t,f} = \mathbf{Q}_f^{\mathsf{H}} \mathbf{z}_{t,f}. \tag{6}$$

The first sub-filter in Eq. (5), which is a multiple-target WPE filter, yields dereverberated sound mixture $\mathbf{z}_{t,f}$ from current observation $\mathbf{x}_{t,f}$ using a prediction matrix $\mathbf{G}_f \in \mathbf{C}^{M(L-D)\times M}$ and a vector containing past observation $\overline{\mathbf{x}}_{t,f} = [\mathbf{x}_{t-D}^\top, \ldots, \mathbf{x}_{t-L+1,f}^\top]^\top \in \mathbb{C}^{M(L-D)\times 1}$. The second sub-filter in Eq. (6) is a separation matrix $\mathbf{Q}_f \in \mathbb{C}^{M\times N}$ that separates dereverberated sound mixture $\mathbf{z}_{t,f} \in \mathbb{C}^{M\times 1}$ into individual sources. The pair of sub-filters (5) and (6) is equivalent to Eq. (4) when they satisfy $\mathbf{Q}_f = \mathbf{W}_{0,f}$ and $\mathbf{G}_f \mathbf{Q}_f = -\left[(\mathbf{W}_{D,f})^\top, \ldots, (\mathbf{W}_{L-1,f})^\top \right]^\top$. This is called source-packed factorization because the sources in the mixture are not distinguished in the output of the WPE filter.

4.2. Optimization with source-packed CBF factorization

Since no closed form solutions are known for the maximization of the likelihood function, conventional techniques utilize iterative estimation based on a coordinate ascent method [19]. It is composed of three estimation steps:

$$\hat{\Theta}_\lambda \leftarrow \underset{\Theta_\lambda}{\operatorname{argmax}}\, \mathcal{L}(\Theta_\lambda, \hat{\Theta}_{\mathbf{Q}}, \hat{\Theta}_{\mathbf{G}}), \tag{7}$$

$$\hat{\Theta}_{\mathbf{Q}} \leftarrow \underset{\Theta_{\mathbf{Q}}}{\operatorname{argmax}}\, \mathcal{L}(\hat{\Theta}_\lambda, \Theta_{\mathbf{Q}}, \hat{\Theta}_{\mathbf{G}}), \tag{8}$$

$$\hat{\Theta}_{\mathbf{G}} \leftarrow \underset{\Theta_{\mathbf{G}}}{\operatorname{argmax}}\, \mathcal{L}(\hat{\Theta}_\lambda, \hat{\Theta}_{\mathbf{Q}}, \Theta_{\mathbf{G}}), \tag{9}$$

where $\Theta_{\mathbf{Q}} = \{\mathbf{Q}_f\}$, $\Theta_{\mathbf{G}} = \{\mathbf{G}_f\}$, and '$\hat{\cdot}$' denotes an estimated variable.

Due to space limitations, we only show the estimation step for $\Theta_{\mathbf{G}}$. Let $\mathbf{g}_{m,f}$ be the mth column of $\mathbf{G}_f$ and $\overline{\mathbf{g}}_f = [\mathbf{g}_{1,f}^\top, \ldots, \mathbf{g}_{M,f}^\top]^\top$. $\Theta_{\mathbf{G}}$ can be updated:

$$\hat{\overline{\mathbf{g}}}_f = \boldsymbol{\Psi}_f^+ \boldsymbol{\psi}_f, \tag{10}$$

$$\boldsymbol{\Psi}_f = \frac{1}{T} \sum_t \overline{\mathbf{X}}_{t,f}^{\mathsf{H}} \boldsymbol{\Phi}_{t,f} \overline{\mathbf{X}}_{t,f} \in \mathbb{C}^{M^2(L-\Delta)\times M^2(L-\Delta)}, \tag{11}$$

$$\boldsymbol{\psi}_f = \frac{1}{T} \sum_t \overline{\mathbf{X}}_{t,f}^{\mathsf{H}} \boldsymbol{\Phi}_{t,f} \mathbf{x}_{t,f} \in \mathbb{C}^{M^2(L-\Delta)\times 1}, \tag{12}$$

where $\overline{\mathbf{X}}_{t,f} = \mathbf{I}_M \otimes \overline{\mathbf{x}}_{t,f}^\top \in \mathbb{C}^{M\times M^2(L-\Delta)}$ letting $\otimes$ denote a Kronecker product, $\boldsymbol{\Phi}_{t,f} = \sum_{n=1}^N \hat{\mathbf{q}}_f^{(n)} \left(\hat{\mathbf{q}}_f^{(n)}\right)^{\mathsf{H}} / \hat{\lambda}_{t,f}^{(n)} \in \mathbb{C}^{M\times M}$ letting $\hat{\mathbf{q}}_f^{(n)}$ be the nth column of $\hat{\mathbf{Q}}_f$, and $(\cdot)^+$ is the Moore-Penrose pseudo-inverse.

In the above update, $\boldsymbol{\Psi}_f$ is the spatio-temporal covariance matrix reflecting the statistics of all the sources. Because $\boldsymbol{\Psi}_f$ is huge, its calculation and its inverse require high computational cost. This is the problem with the conventional method. In the following, we propose techniques to reduce the cost.

5. Proposed optimization method

Two techniques have been proposed to reduce the computational cost of optimizing a mask-based CBF [24]: source-wise covariance decomposition and the source-wise CBF factorization. Below we apply these techniques to blind CBF estimation.

5.1. Source-wise covariance decomposition

By carefully rewriting Eqs. (11) and (12), they can be strictly decomposed into terms for individual sources [24]:

$$\boldsymbol{\Psi}_f = \sum_{n=1}^N \left(\hat{\mathbf{q}}_f^{(n)} \left(\hat{\mathbf{q}}_f^{(n)}\right)^{\mathsf{H}} \otimes \left(\mathbf{R}_{\mathbf{x},f}^{(n)}\right)^\top \right), \tag{13}$$

$$\boldsymbol{\psi}_f = \sum_{n=1}^N \left(\hat{\mathbf{q}}_f^{(n)} \otimes \left(\mathbf{P}_{\mathbf{x},f}^{(n)} \hat{\mathbf{q}}_f^{(n)}\right)^* \right), \tag{14}$$

where $()^*$ denotes a complex conjugate, and $\mathbf{R}_{\mathbf{x},f}^{(n)}$ and $\mathbf{P}_{\mathbf{x},f}^{(n)}$ are respectively a spatio-temporal covariance matrix and a vector of

the nth source:

$$\mathbf{R}_{\mathbf{x},f}^{(n)} = \frac{1}{T}\sum_t \frac{\overline{\mathbf{x}}_{t,f}\overline{\mathbf{x}}_{t,f}^{\mathsf{H}}}{\hat{\lambda}_{t,f}^{(n)}} \in \mathbb{C}^{M(L-D)\times M(L-D)}, \quad (15)$$

$$\mathbf{P}_{\mathbf{x},f}^{(n)} = \frac{1}{T}\sum_t \frac{\overline{\mathbf{x}}_{t,f}\mathbf{x}_{t,f}^{\mathsf{H}}}{\hat{\lambda}_{t,f}^{(n)}} \in \mathbb{C}^{M(L-D)\times M}. \quad (16)$$

In Eqs. (13) and (14), the majority of the calculation is derived from $\mathbf{R}_{\mathbf{x},f}^{(i)}$. Because the matrix is much smaller than $\mathbf{\Psi}_f$, we can greatly reduce the computing cost with this modification in comparison with the direct calculation of Eqs. (11) and (12). Although we still need to calculate the inverse of huge matrix $\mathbf{\Psi}_f$ even with this modification, the cost is relatively small compared with the direct calculation of $\mathbf{\Psi}_f$.

5.2. Source-wise CBF factorization

With source-wise factorization, the CBF in Eq. (4) is first decomposed into a set of CBFs, each of which estimates each source independently:

$$y_{t,f}^{(n)} = \left(\mathbf{w}_{0,f}^{(n)}\right)^{\mathsf{H}}\mathbf{x}_{t,f} + \sum_{\tau=D}^{L-1}\left(\mathbf{w}_{\tau,f}^{(n)}\right)^{\mathsf{H}}\mathbf{x}_{t-\tau,f}, \quad (17)$$

where $\mathbf{w}_{\tau,f}^{(n)} \in \mathbb{C}^{M\times 1}$ for $\tau = 0, D, \ldots, L-1$ is the nth column of $\mathbf{W}_{\tau,f}$. Each CBF is then factorized into two sub-filters:

$$\mathbf{z}_{t,f}^{(n)} = \mathbf{x}_{t,f} - \left(\mathbf{G}_f^{(n)}\right)^{\mathsf{H}}\overline{\mathbf{x}}_{t,f}, \quad (18)$$

$$y_{t,f}^{(n)} = \left(\mathbf{q}_f^{(n)}\right)^{\mathsf{H}}\mathbf{z}_{t,f}^{(n)}. \quad (19)$$

The first sub-filter, which is a single-target WPE filter with a prediction matrix $\mathbf{G}_f^{(n)} \in \mathbf{C}^{M(L-D)\times M}$, dereverberates the nth source. The second sub-filter, which is a beamformer $\mathbf{q}_f^{(n)} \in \mathbb{C}^{M\times N}$, extracts the nth source signal. The pair of sub-filters (18) and (19) is equivalent to Eq. (17) when they satisfy $\mathbf{q}_f^{(n)} = \mathbf{w}_{0,f}^{(n)}$ and $\mathbf{G}_f^{(n)}\mathbf{q}_f^{(n)} = -\left[\left(\mathbf{w}_{D,f}^{(n)}\right)^{\top}, \ldots, \left(\mathbf{w}_{L-1,f}^{(n)}\right)^{\top}\right]^{\top}$. This is called source-wise factorization, which is different from source-packed factorization in that the prediction matrix and the output of the WPE filter are separately estimated for each source.

5.3. Optimization with source-wise CBF factorization

The optimization of the CBFs is conducted based on the coordinate descent method in the same way as source-packed factorization except that we can separately update prediction matrices $\mathbf{G}_f^{(n)}$ for $n = 1, \ldots, N$ one by one.

5.3.1. *Update of* $\Theta_{\mathbf{G}}$

By fixing Θ_λ and $\Theta_{\mathbf{Q}}$ at their previously updated values, the likelihood function for estimating $\Theta_{\mathbf{G}}$ can be rewritten, disregarding constant terms [24]:

$$\mathcal{L}(\Theta_{\mathbf{G}}) = -\sum_{f,n}\left\|\left(\mathbf{G}_f^{(n)} - \left(\mathbf{R}_{\mathbf{x},f}^{(n)}\right)^{-1}\mathbf{P}_{\mathbf{x},f}^{(n)}\right)\hat{\mathbf{q}}_f^{(n)}\right\|_{\mathbf{R}_{\mathbf{x},f}^{(n)}}^2, \quad (20)$$

where $\|\mathbf{x}\|_{\mathbf{R}}^2 = \mathbf{x}^{\mathsf{H}}\mathbf{R}\mathbf{x}$. Interestingly, Eq. (20) can be maximized, not dependent on $\hat{\mathbf{q}}_f^{(n)}$, by updating $\hat{\mathbf{G}}_f^{(n)}$ as

$$\hat{\mathbf{G}}_f^{(n)} \leftarrow \left(\mathbf{R}_{\mathbf{x},f}^{(n)}\right)^{-1}\mathbf{P}_{\mathbf{x},f}^{(n)}. \quad (21)$$

The above update equation is identical to that of the conventional WPE filter optimization, except that $\hat{\lambda}_{t,f}^{(n)}$ is obtained not by the variance of the WPE output but by the beamformer output (section 5.3.2).

In the above equation, since $\mathbf{R}_{\mathbf{x},f}^{(n)}$ can be much smaller than that of $\mathbf{\Psi}_f$, the computational cost for calculating $\mathbf{R}_{\mathbf{x},f}^{(n)}$ and its inverse can be very small. This is an advantage of source-wise factorization over conventional source-packed factorization.

5.3.2. *Update of* $\Theta_{\mathbf{Q}}$ *and* Θ_λ

Methods for updating $\Theta_{\mathbf{Q}}$ and Θ_λ can be derived in the same way as those for the conventional techniques.

To update $\Theta_{\mathbf{Q}}$ and by fixing Θ_λ and $\Theta_{\mathbf{G}}$, the likelihood function can be rewritten, disregarding constant terms:

$$\mathcal{L}(\Theta_{\mathbf{Q}}) = -\sum_{f,n}\left(\mathbf{q}_f^{(n)}\right)^{\mathsf{H}}\Sigma_{\mathbf{z},f}^{(n)}\mathbf{q}_f^{(n)} + 2T\sum_f \log|\det\mathbf{Q}_f|,$$

where $\Sigma_{\mathbf{z},f}^{(n)} = \frac{1}{T}\sum_t \mathbf{z}_{t,f}^{(n)}\left(\mathbf{z}_{t,f}^{(n)}\right)^{\mathsf{H}}/\hat{\lambda}_{t,f}^{(n)}$. According to the idea of Iterative Projection (IP) [5], a solution that maximizes the above function can be obtained by alternately iterating the following updates for each source:

$$\hat{\mathbf{q}}_f^{(n)} \leftarrow \left(\hat{\mathbf{Q}}_f^{\mathsf{H}}\Sigma_{\mathbf{z},f}^{(n)}\right)^{-1}\mathbf{e}_n, \quad (22)$$

$$\hat{\mathbf{q}}_f^{(n)} \leftarrow \left(\left(\hat{\mathbf{q}}_f^{(n)}\right)^{\mathsf{H}}\Sigma_{\mathbf{z},f}^{(n)}\hat{\mathbf{q}}_f^{(n)}\right)^{-1/2}\hat{\mathbf{q}}_f^{(n)}, \quad (23)$$

where $\mathbf{e}_n$ is the nth column of an identity matrix $\mathbf{I}_N \in \mathbb{R}^{N\times N}$.

The likelihood function for the update of Θ_λ, on the other hand, can be rewritten:

$$\mathcal{L}(\Theta_\lambda) = -\sum_{t,f,n}\left(\log\lambda_{t,f}^{(n)} + \frac{|y_{t,f}^{(n)}|^2}{\lambda_{t,f}^{(n)}}\right), \quad (24)$$

and the solution that maximizes the function can be obtained simply as $\hat{\lambda}_{t,f}^{(n)} \leftarrow |y_{t,f}^{(n)}|^2$. This solution, however, contains permutation ambiguity, that is, separation is done independently at each frequency, and the separated sources need to be associated with each other over different frequencies. Several approaches have been proposed to solve this problem [3, 4, 6, 28]. In this paper, we adopt a technique used for Independent Vector Analysis (IVA) through which $\lambda_{t,f}^{(n)}$ is assumed to be independent of the frequency for each source n. With this technique and by dropping the frequency indices from $\lambda_{t,f}^{(n)}$, it is updated:

$$\hat{\lambda}_t^{(n)} \leftarrow \frac{1}{F}\sum_{f=0}^{F-1}|y_{t,f}^{(n)}|^2 \quad (25)$$

where F is the number of frequency bins.

6. EXPERIMENTS

This section experimentally evaluates the performance of our proposed techniques in terms of computational complexity and ASR performance improvement.

6.1. Dataset, evaluation metrics, and methods compared

For our evaluation, we prepared a set of noisy reverberant speech mixtures (REVERB-MIX) using the REVERB Challenge dataset (REVERB) [29]. Each utterance in REVERB

Table 1: *Computing time required for 10 WPE iterations and 100 IVA iterations on a mixture utterance with a 9.44 s length. Computing times were measured on a Matlab interpreter by elapsed time.*

Method	Time (s)
IVA (w/o WPE)	32.5
WPE+IVA (not jointly optimized)	34.0
Conventional CBF (jointly optimized)	306.1
CovDecomp (proposed)	42.9
SWFact (proposed)	44.3

Table 2: *WER (%) of enhanced speech signals for RealData in REVERB-MIX with varying estimation iterations: Without enhancement, WER was 62.49 %.*

Method	#iterations of IVA part				
	40	80	120	160	200
IVA	40.08	38.33	37.41	37.21	37.55
WPE+IVA	32.13	30.21	30.64	30.63	30.98
CovDecomp	31.64	29.97	**29.72**	29.83	29.48
SWFact	**30.96**	**29.59**	**29.72**	**29.63**	**29.15**

contains a single reverberant speech with moderate stationary diffuse noise. For generating a set of test data, we mixed two utterances extracted from REVERB, one from its development set (Dev set) and the other from its evaluation set (Eval set), so that each pair of mixed utterances was recorded in the same room, by the same microphone array and under the same condition (near or far, RealData or SimData). We categorized the test data by the original categories of the data in REVERB (e.g., SimData or RealData). We created an identical number of mixtures in the test data as in the REVERB Eval set, and each utterance in the REVERB Eval set was contained in either one of the mixtures in the test data. Furthermore, the length of each mixture in the test data was set the same as that of the corresponding utterance in the REVERB Eval set.

In the experiments, we estimated three source signals from each mixture, assuming that two of them correspond to the speech signals and the other corresponds to the diffuse noise, and evaluated only one of the speech signals corresponding to the REVERB Eval set. We selected the signal to be evaluated based on the correlation between the estimated signals and the original signal in the REVERB Eval set. As the evaluation metric, we adopted the ASR performance and used a baseline ASR system for REVERB that was developed using Kaldi [30]. This system was composed of a TDNN acoustic model trained using a lattice-free MMI and online i-vector extraction, and a trigram language model. They were trained on the REVERB training set.

We compared our two proposed techniques, i.e., CBF with our proposed covariance decomposition (CovDecomp) and CBF with our proposed source-wise factorization (SWFact), with three conventional methods: CBF with a conventional joint optimization scheme (Conventional CBF), a cascade configuration of WPE followed by IVA (WPE+IVA), and an IVA w/o WPE. For all the methods, we also applied permutation realignment post-processing [6] because it consistently improved the WERs. We set the frame length and the shift at 128 and 32 ms for the IVA and 64 and 16 ms for the other methods based on the settings that achieved the best WERs for the respective methods. A Hann window was used for the short-time analysis. The sampling frequency was 16 kHz and $M = 3$ microphones were used for all the experiments. For the WPE, the prediction delay was set at $D = 2$ and the prediction filter lengths were respectively set at $L = 10, 8$, and 4 for frequency ranges of 0 to 0.8, 0.8 to 1.5, and 1.5 to 8 kHz.

In our preliminary experiments, since WPE converged much faster than IVA within the iterative optimization framework, we set the iteration numbers of WPE 10 times smaller than those of IVA for all the experiments. For example, we updated WPE once every 10 IVA updates in the proposed methods. This iteration scheme is advantageous for making the joint optimization computationally more efficient because the WPE updates are computationally more demanding than those for IVA.

6.2. Evaluation results

Table 1 compares the computing times measured on a Matlab interpreter by the elapsed time for processing a mixture utterance with a 9.44 s length. Both proposed techniques greatly reduced the computing time in comparison with the conventional CBF. When compared with WPE+IVA, the proposed techniques increased the computational cost, although not very much. Note that the difference between IVA (w/o WPE) and WPE+IVA was small because we adopted a longer analysis window for IVA (128 ms) than those for the others (64 ms) so that IVA can achieve its best WER.

Table 2 shows the WERs of the enhanced speech signals obtained using different enhancement methods for RealData in REVERB-MIX. In the experiment, we skipped the evaluation of the conventional CBF because it should be identical to CovDecomp (proposed). As shown in the table, our two proposed methods effectively reduced the WERs in comparison with IVA and WPE+IVA. When we compared the two proposed methods, SWFact slightly outperformed CovDecomp. This is probably because the WPE update does not depend on $\hat{q}_f^{(n)}$ with SWFact (Eq. (21)), which might work advantageously for slightly speeding up the convergence.

7. Concluding remarks

This paper presented two techniques for optimizing a CBF in a computationally efficient way to jointly perform BDR and BSS. One is source-wise covariance decomposition, which provided a computationally efficient way for calculating a huge spatio-temporal covariance matrix that reflects the statistics of all the sources. The other is source-wise CBF factorization, which allows us to skip the calculation of a huge matrix by introducing a new factorization scheme to the CBF optimization. In experiments, we further introduced a computationally efficient iterative optimization scheme, where we updated WPE much less frequently than IVA. Our experiments showed that the proposed techniques greatly reduced the computational cost in comparison with the conventional CBF joint optimization technique without loss of effectiveness.

8. References

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

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley & Sons, 2001.

[3] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. on Speech, and Audio Processing*, vol. 15, no. 1, pp. 70–79, 2006.

[4] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Independent Component Analysis and Blind Signal Separation*, 2006, pp. 601–608.

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

[6] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[7] D. H. T. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE ICASSP*, 2010.

[8] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2010.

[9] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of a dominant source signal from mixtures of many sources," in *Proc. IEEE ICASSP*, vol. III, 2005, pp. 61–64.

[10] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, "Efficient full-rank spatial covariance estimation using independent low-rank matrix analysis for blind source separation," in *Proc. EUSIPCO*, 2019, pp. 1814–1818.

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

[12] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[13] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1509–1520, 2015.

[14] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *Proc. IEEE Int. Workshop on Signal Processing Advances in Wireless Communications*, 1997, pp. 101–107.

[15] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: a versatile framework for multichannel blind signal processing," in *Proc. IEEE ICASSP*, vol. III, 2004, pp. 889–892.

[16] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, January 2011.

[17] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Multichannel speech dereverberation and separation with optimized combination of linear and non-linear filtering," in *Proc. IEEE ICASSP*, 2012, pp. 4057–4060.

[18] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. IWAENC*, 2014.

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

[20] X. Li, L. Girin, S. Gannot, and R. Horaud, "Multichannel source separation and speech enhancement using the convolutive transfer function," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 645–659, March 2019.

[21] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, "Independent low-rank matrix analysis with decorrelation learning," in *Proc. IEEE WASPAA*, October 2019.

[22] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 28, pp. 740–754, January 2020.

[23] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly optimal dereverberation and beamforming," in *Proc. IEEE ICASSP*, 2020.

[24] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2020.

[25] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 903–907, April 2019.

[26] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.

[27] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE trans. Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.

[28] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[29] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, 2016.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011.