TIME-FREQUENCY-MASKING-BASED DETERMINED BSS WITH APPLICATION TO SPARSE IVA

Kohei Yatabe[†] *and Daichi Kitamura*[‡*]

[†]Department of Intermedia Art and Science, Waseda University, Tokyo, Japan
[‡]Department of Electrical and Computer Engineering, National Institute of Technology, Kagawa, Japan

ABSTRACT

Most of the determined blind source separation (BSS) algorithms related to the independent component analysis (ICA) were derived from mathematical models of source signals. However, such derivation restricts the application of algorithms to explicitly definable source models, i.e., an implicit model associated with some signal-processing procedure cannot be utilized within such framework. In this paper, we propose an extension of the existing algorithm so that any time-frequency masking method (e.g., those developed in speech enhancement literature) can be incorporated into the determined BSS algorithm. As an application of the proposed algorithm, a sparse extension of the well-known independent vector analysis (IVA) is also proposed for illustrating the potentiality of the masking-based implicit source model.

Index Terms— Linear source separation, demixing filter estimation, primal-dual splitting, proximity operator, plug-and-play.

1. INTRODUCTION

Blind source separation (BSS) is methodology for recovering source signals from multiple mixtures without any knowledge about the mixing system. Let a convolutive mixing process of audio signals be approximated in time-frequency domain as

$$\mathbf{x}[t,f] \approx A[f]\mathbf{s}[t,f], \tag{1}$$

where $\mathbf{x} = [x_1, x_2, \dots x_M]^T$ is an observation obtained by M microphones, $\mathbf{s} = [s_1, s_2, \dots s_N]^T$ is a source signal to be recovered, $A[f]$ is an $M \times N$ mixing matrix, and t and f are indices of time and frequency, respectively. Then, the aim of BSS is to recover N source signals $\mathbf{s}$ from the mixtures $\mathbf{x}$. In a determined or overdetermined situation ($M \geq N$), many of the BSS problems are formulated as an estimation problem of finding an $N \times M$ demixing matrix $W[f]$ which is a left inverse of $A[f]$ (i.e., $W[f]A[f] = I$), and the source signals are recovered by simple multiplication:

$$W[f]\mathbf{x}[t,f] \approx W[f]A[f]\mathbf{s}[t,f] = \mathbf{s}[t,f]. \tag{2}$$

For the sake of simplicity, only a determined situation ($M = N$) is considered in this paper.

For estimating a demixing matrix $W[f]$, statistical independence between source signals is often assumed, which leads to a family of independence-based BSS algorithms. Arguably, the independent component analysis (ICA) applied in frequency domain (FDICA) [1–3] is one of the most famous methods among them. Some recent developments on this line aim to avoid the so-called permutation problem [4, 5] by considering more sophisticated models of source signals. For instance, the independent vector analysis (IVA) [6–8] assumes co-occurrence among the frequency components in each source, and the independent low-rank matrix analysis (ILRMA) [9–11] assumes low-rankness on spectrogram of each source. These models equipped with the recent algorithms based on the majorization-minimization (MM) principle [8, 11] are often considered as the current state-of-the-art methods.

The key to success of these methods is to incorporate prior knowledge on source signals into their formulations. However, the state-of-the-art MM algorithms may require time for deriving a correct majorizing function tailored for each source model, which delays the trial of a new source model. As an alternative for seeking and developing a better model, a flexible algorithm for determined BSS, based on a proximal splitting method [12–15], has been proposed recently [16] to handle a lot of source models with less effort. It seemed promising since some new models difficult for MM algorithms were shown to be effective [16]. Yet, the algorithm has a fundamental limitation which may prevent a practical use.

To resolve the limitation and widen the application of the BSS algorithm in [16], its heuristic extension is proposed in this paper. The fundamental limitation of the algorithm is its requirement of the proximity operator corresponding to the source model (see Section 2.2). As a proximity operator is an optimization problem, application of the algorithm is limited to the class of source models whose proximity operators can be easily calculated. To weaken this requirement, a time-frequency masking operator is heuristically substituted in place of the proximity operator. As an application of the proposed algorithm, a new BSS model, named sparse IVA, is also proposed to demonstrate the potentiality of the proposed extension.

2. PROXIMAL ALGORITHM FOR DETERMINED BSS

In this section, the proximal algorithm proposed in [16] for the independence-based determined BSS problems is briefly reviewed.

2.1. Independence-based BSS problems

As introduced in the previous section, the aim of determined BSS methods is to estimate $M \times M$ demixing matrices $\{W[f]\}_{f=1}^F$ so that the source signals are approximately recovered from the observations as $W[f]\mathbf{x}[t,f] \approx \mathbf{s}[t,f]$. By assuming statistical independence between the source signals, many of the BSS methods have been formulated as a minimization problem of the following form:

$$\underset{\{W[f]\}_{f=1}^F}{\text{Minimize}} \quad \mathcal{P}(W[f]\mathbf{x}[t,f]) - \sum_{f=1}^{F} \log|\det(W[f])|, \tag{3}$$

where $\mathcal{P}$ is a real-valued penalty function corresponding to the source model. For example, with some constant C, the traditional FDICA, whose source model is the Laplace distribution, is given by

$$\mathcal{P}(\mathbf{y}[t,f]) = C\left\|\mathbf{y}[t,f]\right\|_1 = C\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{f=1}^{F}\left|y_m[t,f]\right|, \tag{4}$$

*This work was partly supported by JSPS KAKENHI Grant-in-Aid for Research Activity Start-up (17H06572, 17H07191).

while IVA based on the spherical Laplace distribution is obtained by

$$\mathcal{P}(\mathbf{y}[t,f]) = C\,\|\mathbf{y}[t,f]\|_{2,1} = C\sum_{m=1}^{M}\sum_{t=1}^{T}\Big(\sum_{f=1}^{F}|y_m[t,f]|^2\Big)^{\frac{1}{2}}. \quad (5)$$

ILRMA can also be interpreted as Eq. (3) in the similar way [16]. From this perspective, it is clear that the difference of performance among these methods is owing to goodness of the penalty function $\mathcal{P}$. Therefore, a BSS method can be improved by finding a better penalty function $\mathcal{P}$ corresponding to a good source model.

2.2. Proximal algorithm for determined BSS problems [16]

As a new BSS method is developed by seeking a better model, it is convenient to have a single algorithm that can handle a lot of models without effort on modifying the code. In [16], the primal-dual splitting (PDS) algorithm [14] was employed to meet this requirement.

By reformulating Eq. (3) into a PDS applicable form,

$$\underset{\mathbf{w}}{\text{Minimize}} \quad \mathcal{I}(\mathbf{w}) + \mathcal{P}(X\mathbf{w}), \quad (6)$$

the proximal algorithm in [16] is obtained as in Algorithm 1, where

$$\mathbf{w} = [\mathbf{w}[1]^T, \mathbf{w}[2]^T, \ldots, \mathbf{w}[F]^T]^T, \quad \big(\,\mathbf{w}[f] = \mathcal{V}(W[f])\,\big) \quad (7)$$

is a vectorized version of the demixing matrices $\{W[f]\}_{f=1}^{F}$, $\mathcal{V}$ is the vectorizing operator converting a matrix into a vector,

$$\mathcal{V}(W[f]) = [W_{1,1}[f], \ldots, W_{1,M}[f], W_{2,1}[f], \ldots, W_{M,M}[f]]^T, \quad (8)$$

$\mathcal{M}$ is the linear operator converting the vector back into the matrix,

$$\mathcal{M}(\mathbf{w})[f] = W[f], \quad (9)$$

X is a matrix constructed from the observed data $\mathbf{x}[t,f]$ as

$$X = \mathrm{blkdiag}(\boldsymbol{\chi}[1], \boldsymbol{\chi}[2], \ldots, \boldsymbol{\chi}[F]), \quad (10)$$

$$\boldsymbol{\chi}[f] = \mathrm{blkdiag}(\chi[f], \chi[f], \ldots, \chi[f]), \quad (M \text{ times}) \quad (11)$$

$$\chi[f] = [\tau_1[f], \tau_2[f], \ldots, \tau_M[f]], \quad (12)$$

$$\tau_m[f] = [x_m[1,f], x_m[2,f], \ldots, x_m[T,f]]^T, \quad (13)$$

$\mathrm{blkdiag}(\cdot)$ is an operator constructing a block-diagonal matrix by concatenating inputted matrices diagonally, $\tau_m[f]$ is $T\times 1$, $\chi[f]$ is $T\times M$, $\boldsymbol{\chi}[f]$ is $MT\times M^2$, X is $FMT\times FM^2$,

$$\mathcal{I}(\mathbf{w}) = -\sum_{f=1}^{F}\sum_{m=1}^{M}\log\sigma_m(\mathcal{M}(\mathbf{w})[f]), \quad (14)$$

and $\sigma_m(W)$ is the mth singular value of W. The step-size parameters can be chosen simply as $\mu_1 = \mu_2 = \alpha = 1$ through the normalization rule (see [16] for details and an extension).

The important feature of this algorithm is that each function in the problem is independently minimized via the proximity operator,

$$\mathrm{prox}_{\mu g}[\mathbf{z}] = \arg\min_{\boldsymbol{\xi}}\Big[\, g(\boldsymbol{\xi}) + \frac{1}{2\mu}\,\|\mathbf{z}-\boldsymbol{\xi}\|_2^2\,\Big], \quad (15)$$

which is a subproblem easier than the original problem [13]. Proximity operators of some functions related to the BSS problems, including $-\log$ in Eq. (3) and norms in Eqs. (4) and (5), can be computed quite efficiently. Therefore, Algorithm 1 is promising because difficulty of the BSS problem only depends on the subproblem of the source model, and a new algorithm is obtained for each source model by only replacing the proximity operator in the 6th line.

Algorithm 1 PDS-BSS [16]

1: **Input:** X, $\mathbf{w}^{[1]}$, $\mathbf{y}^{[1]}$, μ_1, μ_2, α
2: **Output:** $\mathbf{w}^{[K+1]}$
3: **for** $k = 1, \ldots, K$ **do**
4: $\quad\widetilde{\mathbf{w}} = \mathrm{prox}_{\mu_1\mathcal{I}}[\,\mathbf{w}^{[k]} - \mu_1\mu_2 X^H\mathbf{y}^{[k]}\,]$
5: $\quad\mathbf{z} = \mathbf{y}^{[k]} + X(2\widetilde{\mathbf{w}} - \mathbf{w}^{[k]})$
6: $\quad\widetilde{\mathbf{y}} = \mathbf{z} - \mathrm{prox}_{\frac{1}{\mu_2}\mathcal{P}}[\,\mathbf{z}\,]$
7: $\quad\mathbf{y}^{[k+1]} = \alpha\widetilde{\mathbf{y}} + (1-\alpha)\mathbf{y}^{[k]}$
8: $\quad\mathbf{w}^{[k+1]} = \alpha\widetilde{\mathbf{w}} + (1-\alpha)\mathbf{w}^{[k]}$
9: **end for**

Although proximity operators divide the problem into simpler subproblems, they are still optimization problems whose solutions may not be obtained easily. That is, Algorithm 1 is limited to the class of source models whose proximity operators are easily computable. Moreover, for deriving a solution to the proximity operator, the penalty function of the source model must be written explicitly, which prohibits a use of implicit source models learned from data.

3. PROPOSED METHOD

To overcome the limitation and widen the application of Algorithm 1, the proximity operator is heuristically replaced by a time-frequency masking operator. To do so, the connection between the proximity operator and time-frequency masking is discussed first.

3.1. Proximity operators as time-frequency masking

Some source models admit closed-form solutions to the associated proximity operators. For example, the proximity operator of ℓ_1 norm in Eq. (4) is given by the bin-wise soft-thresholding operator,

$$\big(\mathrm{prox}_{\lambda\|\cdot\|_1}[\mathbf{z}]\big)_m[t,f] = \left(1 - \frac{\lambda}{|z_m[t,f]|}\right)_+ z_m[t,f], \quad (16)$$

where $(\cdot)_+ = \max\{0,\cdot\}$ is the projection onto nonnegative values, and $\lambda \geq 0$. That of $\ell_{2,1}$ mixed norm in Eq. (5) is also given as

$$\big(\mathrm{prox}_{\lambda\|\cdot\|_{2,1}}[\mathbf{z}]\big)_m[t,f] = \left(1 - \frac{\lambda}{(\sum_{f=1}^{F}|z_m[t,f]|^2)^{\frac{1}{2}}}\right)_+ z_m[t,f], \quad (17)$$

which is called the group-thresholding operator. By inserting one of these operators into the 6th line, Algorithm 1 for FDICA or IVA is obtained. Proximity operators of many other sparsity-inducing functions can also be obtained as thresholding operators [17, 18].

Note that the above proximity operators have the same form:

$$\big(\mathcal{T}_\lambda[\mathbf{z}]\big)_m[t,f] = \big(\mathcal{M}(\mathbf{z})\big)_m[t,f]\; z_m[t,f], \quad (18)$$

where $0 \leq (\mathcal{M}(\mathbf{z}))_m[t,f] \leq 1$ is a scalar depending on the input, and the soft- and group-thresholding operators in Eqs. (16) and (17) are obtained by inserting the following functions into Eq. (18):

$$\big(\mathcal{M}^{\lambda}_{\ell_1}(\mathbf{z})\big)_m[t,f] = \big(1 - \lambda/|z_m[t,f]|\big)_+, \quad (19)$$

$$\big(\mathcal{M}^{\lambda}_{\ell_{2,1}}(\mathbf{z})\big)_m[t,f] = \big(1 - \lambda/(\textstyle\sum_{f=1}^{F}|z_m[t,f]|^2)^{\frac{1}{2}}\big)_+. \quad (20)$$

This form can be interpreted as the time-frequency masking whose mask $\mathcal{M}(\mathbf{z})$ is given by a procedure depending on its input.

3.2. Proximity operator as MAP estimation

The optimization problem of the proximity operator in Eq. (15) can be interpreted as the maximum *a posteriori* (MAP) estimation. When an observed signal is contaminated by the additive Gaussian noise, MAP estimation of the clean signal, whose prior distribution is $C\exp(-\mathcal{P}(\cdot))$, reduces to the following maximization problem:

$$\mathrm{prox}_{\mu\mathcal{P}}[\mathbf{z}] = \arg\max_{\boldsymbol{\xi}} \left[e^{-\frac{1}{2\mu}\|\mathbf{z}-\boldsymbol{\xi}\|_2^2} e^{-\mathcal{P}(\boldsymbol{\xi})} \right], \qquad (21)$$

which is equivalent to the proximity operator (taking negative logarithm of Eq. (21) recovers Eq. (15)). This interpretation suggests that substituting a general Gaussian denoiser, which (approximately) solves Eq. (21), in place of the proximity operator results in an algorithm which works as if the penalty function $\mathcal{P}$ is minimized. Such idea of replacing a proximity operator is called plug-and-play method [19] which is attracting many researchers recently [20–23].

3.3. Proposed algorithm

Based on the above relations, the proximity operator in Algorithm 1 is replaced by a time-frequency mask as in Algorithm 2, where $\odot$ denotes the element-wise product, and θ represents a set of parameters for generating the mask. This slight generalization allows collaboration of time-frequency masking and the determined BSS algorithm. It greatly extends the possibility because explicit form of the penalty function $\mathcal{P}$ is not required, i.e., any mask (maybe defined as a rule and/or learned from data) can be incorporated into the algorithm.

When the underlying penalty function is separable for each source as in Section 2.1 ($\mathcal{P} = \sum_{n=1}^{N} \mathcal{P}_n$), then the proposed algorithm can be seen as an independence-based BSS method (ML estimation) with $C\exp(-\mathcal{P}_n(\cdot))$ being the density function of nth source signal. That is, the proposed algorithm recasts the BSS problem in Eq. (3) into the denoising problem in Eq. (21) consisting of the same prior distribution of the sources. This is important property because learning a Gaussian denoiser is much easier than learning a regressor of the demixing matrix which requires a variety of impulse responses as the training data. It is also possible to obtain an algorithm beyond the independence-based framework by inserting a masking method which is not separable for each source.

In practice, one can insert any time-frequency mask into Algorithm 2 to generate a new BSS algorithm. Although stability and convergence of the algorithm for a general time-frequency mask can only be investigated by experiments, it is easy since the only effort for rewriting the code is the masking function in the 6th line of Algorithm 2. That is, one can just insert a masking method into the algorithm and run it for checking the performance.

4. SPARSE IVA: SIMPLE YET POWERFUL EXTENSION OF IVA

As an application of the proposed algorithm, a computationally cheap but well-performing extension of IVA, named Sparse IVA, is proposed here. Although it might seem more natural to consider a recent learning-based masking method (such as ones using deep neural networks) from the above discussion, our motivation is to show that there is still a room for improving the well-known model for which the proposed concept is essential.

4.1. Side effect of whitening undesirable for IVA

As for many other algorithms, whitening is strongly recommended before running the algorithm. However, it causes a side effect which degrades the performance of IVA. Since the frequency-wise energy of the observed data is set to the same value for every frequencies, whitening distorts the assumption of IVA, co-occurrence among the frequency components, which prevent IVA from working properly.

Algorithm 2 PDS-BSS-masking

1: **Input:** X, $\mathbf{w}^{[1]}$, $\mathbf{y}^{[1]}$, μ_1, μ_2, α
2: **Output:** $\mathbf{w}^{[K+1]}$
3: **for** $k = 1, \ldots, K$ **do**
4: $\quad \widetilde{\mathbf{w}} = \mathrm{prox}_{\mu_1 \mathcal{I}}[\, \mathbf{w}^{[k]} - \mu_1\mu_2 X^H \mathbf{y}^{[k]} \,]$
5: $\quad \mathbf{z} = \mathbf{y}^{[k]} + X(2\widetilde{\mathbf{w}} - \mathbf{w}^{[k]})$
6: $\quad \widetilde{\mathbf{y}} = \mathbf{z} - \mathcal{M}^{\theta}(\mathbf{z}) \odot \mathbf{z}$
7: $\quad \mathbf{y}^{[k+1]} = \alpha\widetilde{\mathbf{y}} + (1-\alpha)\mathbf{y}^{[k]}$
8: $\quad \mathbf{w}^{[k+1]} = \alpha\widetilde{\mathbf{w}} + (1-\alpha)\mathbf{w}^{[k]}$
9: **end for**

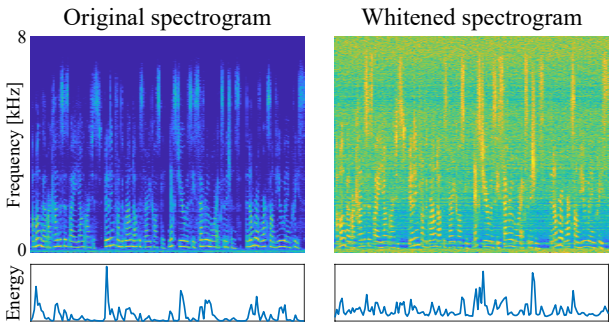

Fig. 1. Illustration of the undesired side effect of whitening. Time-frame-wise energies of each spectrogram are shown below.

To see this effect, Fig. 1 illustrates an example of spectrograms before and after whitening. As seen in the figure, the original speech mixture does not contain very low- and high-frequency components, say, below 100 Hz and above 6 kHz, respectively. These low- and high-frequency bands, which are dominated by ambient noise, are also normalized to have the same energy as other bands. Such noise magnification distorts the group sparse nature of speech signals, and therefore the performance of IVA is degraded. The group sparse structure should be recovered without harming the positive effect of whitening such as improvement on convergence speed.

4.2. Proposed Sparse IVA

By adding ℓ_1 norm to $\ell_{2,1}$ mixed norm, a new BSS model, namely a sparsely regularized IVA, can be derived. By just inserting [24]

$$\mathrm{prox}_{\lambda_1\|\cdot\|_{2,1}+\lambda_2\|\cdot\|_1}[\mathbf{z}] = \mathrm{prox}_{\lambda_1\|\cdot\|_{2,1}}[\,\mathrm{prox}_{\lambda_2\|\cdot\|_1}[\mathbf{z}]\,] \qquad (22)$$

to the 6th line of Algorithm 1, the algorithm for such sparse variant of IVA is obtained [16]. Starting from this composite thresholder, the proposed Sparse IVA is defined through a time-frequency mask.

Firstly, for recovering the group sparse structure distorted by whitening, a frequency-wise weight $(\Theta_\eta[\mathbf{x}])_f \geq 0$ is proposed based on sparseness of each frequency band measured by the normalized ℓ_1 norm. As ℓ_1 norm takes a small value for a sparse signal, we consider the reciprocal of the normalized ℓ_1 norm [25] for the weight:

$$\Theta_\eta[\mathbf{x}] = \Upsilon_\eta\left[\left(\sum_{m=1}^{M}\sum_{t=1}^{T}|x_m[t,f]|^2\right)^{\frac{1}{2}} \Big/ \left(\sum_{m=1}^{M}\sum_{t=1}^{T}|x_m[t,f]|\right)\right], \qquad (23)$$

where $\Upsilon_\eta[\,\cdot\,]$ denotes ℓ_1 normalization with clipping by $\eta \geq 0$,

$$\Upsilon_\eta[\boldsymbol{\xi}] = \boldsymbol{\xi}_\eta / (\|\boldsymbol{\xi}_\eta\|_1 / F), \qquad \boldsymbol{\xi}_\eta = (\boldsymbol{\xi} - \eta)_+, \tag{24}$$

and division in Eq. (23) is performed element-wise. The fth element $(\Theta_\eta[\mathbf{x}])_f$ is larger when the corresponding frequency band is sparser, and vice versa. Clipping by η enforces elements less than η to zero for further suppressing the effect of noisy frequency bands.

Furthermore, enhanced thresholders are employed to reduce the bias imposed by the soft- and group-thresholding. While these thresholding functions in Section 3.1 were obtained from the penalty functions, it is possible to define a thresholder without defining the associated penalty function. Some recent research considers this way to realize a better thresholding function [26–30]. For example, the thresholding rule obtained through the p-shrinkage [27],

$$\big(\mathcal{T}_{p,\lambda}[\mathbf{z}]\big)_m[t,f] = \big(1 - \lambda^{2-p} / |z_m[t,f]|^{2-p}\big)_+ z_m[t,f], \tag{25}$$

is related to a penalty function which does not have an explicit formula for general p. Nevertheless, it behaves as a reasonable thresholding function since $p = 1$ results in the soft-thresholding in Eq. (16), and $p \to -\infty$ corresponds to the hard-thresholding. Another example is one of the social sparsity operators [26],

$$\big(\mathcal{T}_{h,\lambda}[\mathbf{z}]\big)_m[t,f] = \big(1 - \lambda / \sqrt{h * |z_m[t,f]|^2}\big)_+ z_m[t,f], \tag{26}$$

where $h*$ represents convolution with a two-dimensional filter kernel h in time-frequency domain. Although Eq. (26) is not a proximity operator in general, its effectiveness is empirically known as in [26]. In this paper, the firm thresholder [31, 32] is utilized for simplicity.

Then, by imposing the frequency-wise weight $\Theta_\eta[\mathbf{x}]$ and the debiasing operator $\Xi_\kappa[\cdot]$ corresponding to the firm thresholder into Eq. (22), the time-frequency mask for Sparse IVA is proposed:

$$\big(\mathcal{M}^{\mathbf{x},\eta,\boldsymbol{\lambda},\kappa}_{\text{SparseIVA}}(\mathbf{z})\big)_m[t,f] = \tag{27}$$
$$\Xi_\kappa\Bigg[\bigg(1 - \frac{\lambda_1}{\big(\sum_{f=1}^F (\Theta_\eta[\mathbf{x}])_f \, |\zeta^{\mathbf{z},\kappa}_m[t,f]\, z_m[t,f]|^2\big)^{\frac{1}{2}}}\bigg)_+\Bigg] \zeta^{\mathbf{z},\kappa}_m[t,f],$$

where $(\Theta_\eta[\mathbf{x}])_f \geq 0$ is the fth element of $\Theta_\eta[\mathbf{x}]$ in Eq. (23), $\eta \geq 0$ is the clipping parameter in Eq. (24), $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]$ is the thresholds,

$$\big(\Xi_\kappa[\mathbf{z}]\big)_m[t,f] = \big(\kappa\, z_m[t,f] / \max_{m,t,f}\{z_m[t,f]\}\big)_- \tag{28}$$

is a debiasing operator with a magnification factor $\kappa \geq 1$, $(\cdot)_- = \min\{1, \cdot\}$ is a clipping operator so that $\big((z)_+\big)_- \in [0,1]$, and

$$\zeta^{\mathbf{z},\kappa}_m[t,f] = \Xi_\kappa\big[\big(1 - \lambda_2 / |z_m[t,f]|\big)_+\big] \tag{29}$$

is an element-wise mask corresponding to the firm thresholding whose threshold varies depending on the maximum value of the input. This mask can be regarded as a weighted and non-convex version of the thresholding function in Eq. (22) which is recovered by $\kappa = \max_{m,t,f}\{z_m[t,f]\}$ and $(\Theta_\eta[\mathbf{x}])_f = 1$ (for all f).

By inserting this time-frequency mask into Algorithm 2, the BSS algorithm for Sparse IVA is obtained. Although it might seem complicated, calculation of the proposed mask is not so expensive than that of the ordinary Laplace IVA in Eq. (20) since the computational cost of this algorithm is dominated by the 4th and 5th lines involving the (demix) filtering. Note that the proposed algorithm is essential for Sparse IVA because it is defined only by the mask-generating operator and its penalty function is not explicit.

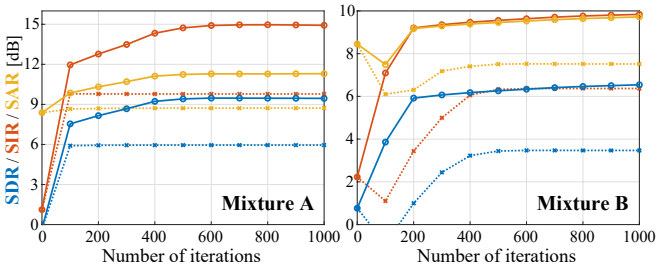

Fig. 2. Comparison between the ordinary IVA based on Eq. (20) (dotted lines) and the proposed Sparse IVA in Eq. (27) (solid lines).

Table 1. Scores at the last iteration of Fig. 2. Run time per iteration was measured by Core i5-7200U processor and MATLAB 2017a.

	Mixture A			Mixture B			Run time
	SDR	SIR	SAR	SDR	SIR	SAR	[ms / iter.]
IVA	6.0	9.8	8.7	3.4	6.3	7.5	55.2
Sparse IVA	9.5	14.9	11.3	6.5	9.8	9.7	67.1
Difference	3.5	5.1	2.6	3.1	3.5	2.2	11.9
Ratio	1.6 x	1.5 x	1.3 x	1.9 x	1.6 x	1.3 x	1.2 x

5. EXPERIMENT

The proposed Sparse IVA was tested by applying it to speech mixtures as in [16]. The database used in this experiment was a part of SiSEC (`dev1` in the UND task[1]). Live recording (`liverec`) of four female speech sources recorded by two microphones (5 cm spacing) was chosen as the test data. For making the problem determined, two pairs of sources were considered: **Mixture A** consists of two sources arrived from -50° and 45° and **Mixture B** consists of two sources arrived from -10° and 15°, where 0° corresponds to the normal direction to the microphone array. The reverberation time was 130 ms, and 128-ms-long Hann window with 64-ms shift was used. The initial value of demixing matrices $\mathbf{w}^{[1]}$ was set to the identity matrices ($W[f] = I$ for all f), and that of $\mathbf{y}$ was the zero vector. The parameters for Sparse IVA were set to $\mu_1 = 1$, $\mu_2 = 1$, $\alpha = 1.75$, $\lambda_1 = 2$, $\lambda_2 = 0.01$, $\kappa = 1.1$, and $\eta = 0.5$. For comparison, the ordinary Laplace IVA based on Eq. (20) was also performed by the proposed algorithm with the same parameters.

The experimental results are shown in Fig. 2 whose scores at the last iteration are listed in Table 1. While the proposed Sparse IVA involved the computational load similar to the ordinary IVA, their scores greatly differ as in the table (SDR improved 3.3 dB in average by only requiring 1.2 x computational efforts). This result indicates that a better BSS algorithm can be obtained with the proposed method by simply designing a mask-generation rule based on a prior knowledge which can be borrowed from existing methods [33–35].

6. CONCLUSIONS

In this paper, a general BSS algorithm which can be easily defined by a time-frequency masking rule was proposed. The proposed method can easily unify a sound enhancement method into determined BSS by simply inserting a mask-generating function into the single line of the algorithm. Its extraordinary flexibility opened up a new frontier of determined BSS since any denoising method based on time-frequency masking can be combined into the BSS algorithm, which was demonstrated through the proposed Sparse IVA.

[1] Available at `http://sisec2011.wiki.irisa.fr`

7. REFERENCES

[1] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[2] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[3] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 666–678, Mar. 2006.

[4] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1, pp. 1–24, 2001.

[5] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

[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, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, Jan. 2007.

[8] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2011, pp. 189–192.

[9] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 276–280.

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," in *Proc. Eur. Signal Process. Conf.*, Aug. 2015, pp. 1261–1265.

[11] 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, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

[12] P. L. Combettes and J.-C. Pesquet, *Proximal Splitting Methods in Signal Processing*, pp. 185–212, Springer, 2011.

[13] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.

[14] N. Komodakis and J. C. Pesquet, "Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 31–54, Nov. 2015.

[15] M. Burger, A. Sawatzky, and G. Steidl, *First Order Algorithms in Variational Image Processing*, pp. 345–407, Springer, 2016.

[16] K. Yatabe and D. Kitamura, "Determined blind source separation via proximal splitting algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 776–780.

[17] M. Kowalski, "Sparse regression using mixed norms," *Appl. Comput. Harm. Anal.*, vol. 27, no. 3, pp. 303–324, 2009.

[18] T. Tachikawa, K. Yatabe, and Y. Oikawa, "Underdetermined source separation with simultaneous DOA estimation without initial value dependency," in *Proc. Int. Workshop Acoust. Signal Enhanc.*, Sep. 2018, pp. 161–165.

[19] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *IEEE Glob. Conf. Signal Inf. Process.*, Dec. 2013, pp. 945–948.

[20] S. H. Chan, X. Wang, and O. A. Elgendy, "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications," *IEEE Trans. Comput. Imaging*, vol. 3, no. 1, pp. 84–98, Mar. 2017.

[21] S. Ono, "Primal-dual plug-and-play image restoration," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1108–1112, Aug. 2017.

[22] U. S. Kamilov, H. Mansour, and B. Wohlberg, "A plug-and-play priors approach for solving nonlinear imaging inverse problems," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1872–1876, Dec. 2017.

[23] T. Meinhardt, M. Moeller, C. Hazirbas, and D. Cremers, "Learning proximal operators: Using denoising networks for regularizing inverse imaging problems," in *IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1799–1808.

[24] A. Gramfort, D. Strohmeier, J. Haueisen, M.S. Hamalainen, and M. Kowalski, "Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations," *NeuroImage*, vol. 70, pp. 410–422, 2013.

[25] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Dec. 2004.

[26] M. Kowalski, K. Siedenburg, and M. Dorfler, "Social sparsity! Neighborhood systems enrich structured shrinkage operators," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2498–2511, May 2013.

[27] R. Chartrand, "Shrinkage mappings and their induced penalty functions," in *IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 1026–1029.

[28] M. Kowalski, "Thresholding RULES and iterative shrinkage/thresholding algorithm: A convergence study," in *IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 4151–4155.

[29] I. W. Selesnick and I. Bayram, "Sparse signal estimation by maximally sparse convex optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1078–1092, Mar. 2014.

[30] I. Bayram, "Penalty functions derived from monotone mappings," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 265–269, Mar. 2015.

[31] H.-Y. Gao and A. G. Bruce, "Waveshrink with firm shrinkage," *Statistica Sinica*, vol. 7, no. 4, pp. 855–874, 1997.

[32] A. Antoniadis, "Wavelet methods in statistics: some recent developments and their applications," *Statist. Surv.*, vol. 1, pp. 16–55, 2007.

[33] K. Yatabe and Y. Oikawa, "Phase corrected total variation for audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 656–660.

[34] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Phase-aware harmonic/percussive source separation via convex optimization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019.

[35] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Data-driven design of perfect reconstruction filterbank for DNN-based sound source enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019.