LINEAR MULTICHANNEL BLIND SOURCE SEPARATION BASED ON TIME-FREQUENCY MASK OBTAINED BY HARMONIC/PERCUSSIVE SOUND SEPARATION

Soichiro Oyabu[†]*, Daichi Kitamura*[†]*, Kohei Yatabe*[‡]

[†]National Institute of Technology, Kagawa College, Kagawa, Japan
[‡]Waseda University, Tokyo, Japan

ABSTRACT

Determined blind source separation (BSS) extracts the source signals by linear multichannel filtering. Its performance depends on the accuracy of source modeling, and hence existing BSS methods have proposed several source models. Recently, a new determined BSS algorithm that incorporates a time-frequency mask has been proposed. It enables very flexible source modeling because the model is implicitly defined by a mask-generating function. Building up on this framework, in this paper, we propose a unification of determined BSS and harmonic/percussive sound separation (HPSS). HPSS is an important preprocessing for musical applications. By incorporating HPSS, both harmonic and percussive instruments can be accurately modeled for determined BSS. The resultant algorithm estimates the demixing filter using the information obtained by an HPSS method. We also propose a stabilization method that is essential for the proposed algorithm. Our experiments showed that the proposed method outperformed both HPSS and determined BSS methods including independent low-rank matrix analysis.

***Index Terms*—** Determined blind source separation (BSS), harmonic/percussive sound separation (HPSS), time-frequency masking, mask stabilization, plug-and-play scheme.

1. INTRODUCTION

Blind source separation (BSS) is a technique for separating individual sources from an observed mixture without any prior knowledge on the mixing system such as the positions of microphones and source signals. In the (over-)determined situation (the number of microphones is greater than or equal to the number of sources), several determined BSS algorithms have been developed based on the assumption of statistical independence between the sources, e.g., the frequency-domain independent component analysis (FDICA) [1–3], the independent vector analysis (IVA) [4–6], and the independent low-rank matrix analysis (ILRMA) [7, 8]. This paper focuses on such a determined BSS algorithm.

Recently, a new class of determined BSS algorithms, which we call the time-frequency-masking-based BSS (TFMBSS), has been proposed [9]. Since source modeling is the key to success, TFMBSS aims at integrating a source model that has not been utilized in determined BSS. To give some examples of conventional source modeling, IVA models co-occurrence among the frequency components of each source, and ILRMA models co-occurrence among the components via the low-rank time-frequency structure of the source signals. These BSS methods have proposed such sophisticated source models because the performance of BSS can be improved by a better model. TFMBSS generalized a source model to a general time-frequency mask so that it extends the class of source models to a wider one. Some new models have successfully obtained well-performing BSS algorithms using TFMBSS [9, 10], and this strategy has capability of further improvement by discovering a better model. In particular, an application-specific source model should be promising.

Harmonic/percussive sound separation (HPSS) [11–15] is one application that requires specific modeling of source signals. Harmonic and percussive instruments have very different roles in music, and hence separating them is crucial for many applications including music analysis (estimation of chords, tempo, rhythm, notes, genre, etc.) and remixing. HPSS separates them by modeling their distinct spectral structures (i.e., smoothness of magnitude spectrograms along time or frequency axes). This is a concept that does not exist in the conventional determined BSS methods handling all source signals by the same source model. Combination of HPSS and determined BSS should be possible to realize a high-quality HPSS method for multichannel observation because determined BSS performs linear filtering that do not cause nonlinear distortion. However, such combination has not been proposed yet owing to the difficulty of integrating the source models of HPSS into the conventional BSS methods.

In this paper, by taking full advantage of the modeling capability of TFMBSS, we propose a multichannel HPSS method for spatially mixed harmonic and percussive instruments. The proposed method incorporates a single-channel HPSS method into the iteration of TFMBSS so that the demixing filter is informed by the HPSS method. A stabilization technique that is essential for the proposed method is also developed in order to safely update the parameters. By the experiments, it was shown that the proposed method outperformed both HPSS and determined BSS methods.

2. PRELIMINARIES

2.1. Harmonic/percussive sound separation (HPSS)

HPSS [11–15] is a method for separating harmonic and percussive sources based on the source models adapted for their distinct spectral patterns: the spectrograms of harmonic and percussive sources have time- and frequency-continuous structures, respectively. Let $\boldsymbol{B} \in \mathbb{C}^{I\times J}$ be the complex-valued spectrogram of the observed signal. From this monaural mixture, HPSS in [12, 13] separates the harmonic and percussive components, $\boldsymbol{H} \in \mathbb{C}^{I\times J}$ and $\boldsymbol{P} \in \mathbb{C}^{I\times J}$, respectively, by solving the following optimization problem:

$$\min_{\boldsymbol{H},\boldsymbol{P}} \ \mathcal{J}(\boldsymbol{H},\boldsymbol{P}) \ \text{ s.t. } \ |\boldsymbol{B}|^{\xi} = |\boldsymbol{H}|^{\xi} + |\boldsymbol{P}|^{\xi}, \tag{1}$$

where $|\cdot|$ represents the element-wise absolute value,

$$\mathcal{J}(\boldsymbol{H},\boldsymbol{P}) = \sum_{i,j} \Big[\kappa_H \left(|h_{i(j+1)}|^{\rho} - |h_{ij}|^{\rho}\right)^2 + \kappa_P \left(|p_{(i+1)j}|^{\rho} - |p_{ij}|^{\rho}\right)^2\Big], \tag{2}$$

This work was partly supported by JSPS KAKENHI under Grants 19K20306 and 19H01116.

ξ and ρ are the domain parameters, h_{ij} and p_{ij} are the elements of $\boldsymbol{H}$ and $\boldsymbol{P}$, respectively, and $\kappa_H > 0$ and $\kappa_P > 0$ are the weight coefficients for each term. The experiment in [13] confirmed that $\xi = 2\rho = 1$ provides a better separation performance. For $\xi = 2\rho$, (1) can be minimized by iterating the following update rules: [12,13]

$$|h_{ij}|^{2\rho} = \frac{c_{ij}|b_{ij}|^{2\rho}}{c_{ij}+d_{ij}}, \qquad |p_{ij}|^{2\rho} = \frac{d_{ij}|b_{ij}|^{2\rho}}{c_{ij}+d_{ij}}, \tag{3}$$

$$c_{ij} = \kappa_H^2 \left(|h_{(i+1)j}|^{\rho} + |h_{(i-1)j}|^{\rho}\right)^2, \tag{4}$$

$$d_{ij} = \kappa_P^2 \left(|p_{i(j+1)}|^{\rho} + |p_{i(j-1)}|^{\rho}\right)^2, \tag{5}$$

where b_{ij} is the element of $\boldsymbol{B}$.

2.2. Determined blind source separation (BSS)

The determined BSS aims at separating each source from an observed multichannel mixture. Let the numbers of sources and microphones be denoted by N and M, respectively. The source, observed, and estimated (separated) signals obtained via the short-time Fourier transform (STFT) are respectively denoted as

$$\boldsymbol{s}_{ij} = [s_{ij1}, s_{ij2}, \cdots, s_{ijn}, \cdots, s_{ijN}]^{\mathrm{T}} \in \mathbb{C}^N, \tag{6}$$

$$\boldsymbol{x}_{ij} = [x_{ij1}, x_{ij2}, \cdots, x_{ijm}, \cdots, x_{ijM}]^{\mathrm{T}} \in \mathbb{C}^M, \tag{7}$$

$$\boldsymbol{y}_{ij} = [y_{ij1}, y_{ij2}, \cdots, y_{ijn}, \cdots, y_{ijN}]^{\mathrm{T}} \in \mathbb{C}^N, \tag{8}$$

where $i = 1, 2, \cdots, I$, $j = 1, 2, \cdots, J$, $n = 1, 2, \cdots, N$, and $m = 1, 2, \cdots, M$ are the indices of frequency bins, time frames, sources, and microphones (channels), respectively, and $\cdot^{\mathrm{T}}$ denotes the transpose. We also denote the complex-valued spectrogram of the mth observed signals in (7) as $\boldsymbol{X}_m \in \mathbb{C}^{I\times J}$, whose elements are x_{ijm}.

We assume the linear time-invariant mixing system as

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}, \tag{9}$$

where $\boldsymbol{A}_i \in \mathbb{C}^{M\times N}$ is a frequency-wise mixing matrix, which depends on the mixing condition, such as locations of sources and microphones. This mixing model holds when the reverberation time is sufficiently shorter than the window length used in STFT. If $M = N$ and $\boldsymbol{A}_i$ is invertible, the estimated signals can be obtained by multiplying a demixing matrix $\boldsymbol{W}_i \approx \boldsymbol{A}_i^{-1} \in \mathbb{C}^{N\times M}$ as

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}. \tag{10}$$

Since (10) is equivalent to linear filtering, artificial distortion in the estimated signals $\boldsymbol{y}_{ij}$ are minimized compared to those obtained by nonlinear separation techniques, such as deep neural networks and single-channel source separation. This is a strong motivation of applying multichannel determined BSS. For this reason, in this paper, we only focus on the determined case $M = N = 2$ (harmonic and percussive sources). For the case $M > N$, a dimensionality reduction method can be utilized to set $M = N$.

2.3. Time-frequency-masking-based BSS (TFMBSS)

To blindly estimate a demixing matrix $\boldsymbol{W}_i$ in (10), TFMBSS has been proposed in [9] as a new determined BSS framework. It generalized a BSS algorithm in [16] that is based on a proximal splitting algorithm [17–20]. The important feature of TFMBSS is that the source model for determined BSS is implicitly defined via a time-frequency mask. Therefore, this algorithm enables collaboration of determined BSS with a time-frequency masking method.

Algorithm 1 TFMBSS [9, 10]

Input: X, $\mathbf{w}^{[1]}$, $\mathbf{y}^{[1]}$, μ_1, μ_2, α
Output: $\mathbf{w}^{[K+1]}$

1: **for** $k = 1, \cdots, K$ **do**
2: $\quad\widetilde{\mathbf{w}} = \mathrm{prox}_{\mu_1 \mathcal{I}}\left[\mathbf{w}^{[k]} - \mu_1\mu_2 X^{\mathrm{H}}\mathbf{y}^{[k]}\right]$
3: $\quad\mathbf{z} = \mathbf{y}^{[k]} + X(2\widetilde{\mathbf{w}} - \mathbf{w}^{[k]})$
4: $\quad\mathcal{M} = \texttt{generateMask}(\mathbf{z})$
5: $\quad\widetilde{\mathbf{y}} = \mathbf{z} - \mathcal{M}\odot\mathbf{z}$
6: $\quad\mathbf{y}^{[k+1]} = \alpha\widetilde{\mathbf{y}} + (1-\alpha)\mathbf{y}^{[k]}$
7: $\quad\mathbf{w}^{[k+1]} = \alpha\widetilde{\mathbf{w}} + (1-\alpha)\mathbf{w}^{[k]}$
8: **end for**

Let $\mathbf{w}$ be a vectorized form of the demixing matrices $(\boldsymbol{W}_i)_{i=1}^I$, and X be the corresponding matrix composed of the observed spectrograms $(\boldsymbol{X}_m)_{m=1}^M$, so that, for all i, j, (10) can be compactly written as $\mathbf{y} = X\mathbf{w}$. By this vectorization and matricization, TFMBSS is given as in Algorithm 1, where $\odot$ is the element-wise product, $\mathrm{prox}_{\mu_1\mathcal{I}}$ is the singular-value operation defined in [16], and μ_1, μ_2 and α are easily determined step-size parameters (see [10] for details of the algorithm). The function $\texttt{generateMask}(\cdot)$ in the fourth line generates a time-frequency mask $\mathcal{M}$. When this mask enhances the source signals, the TFMBSS algorithm works as if the enhanced signals are obtained by the source models utilized for determined BSS. In other words, the source model of determined BSS is implicitly defined through the mask. Any mask aiming at the target signals can be inserted into TFMBSS, and hence the problem of source modeling is reduced to design of a mask-generating function.

3. PROPOSED METHOD

3.1. Motivation

The single-channel HPSS as in Section 2.1 is frequently used as a preprocessing for music analysis. However, due to the nonlinear separation mechanism, the separated harmonic and percussive sounds are contaminated by artificial distortions, e.g., musical noise. Such distortions can be harmful for the following process and degrade the artistic value of music signals. In contrast, the linear spatial separation (10) can minimize the distortion in the separated signals. Hence, combination with determined BSS should be valuable for HPSS.

To achieve distortion-less separation of harmonic and percussive instruments, we propose a unification of single-channel HPSS and determined BSS. The proposed method is based on TFMBSS whose source model is implicitly defined via HPSS. At each iteration of TFMBSS, HPSS is performed for updating the mask $\mathcal{M}$. Then, TFMBSS updates the demixing filter $\mathbf{w}$ based on the mask. To reduce instability of mask generation, we also propose a mask-smoothing process that is essential for the proposed method. This smoothing process can stabilize the update of demixing filter, resulting in better estimation of the separated signals.

3.2. Determined BSS algorithm based on HPSS

A block diagram of the proposed algorithm is shown in Fig. 1. As usual in determined BSS, the observed signal is firstly converted by STFT. Then, a BSS algorithm is applied to estimate the demixing filter. After separation by (10), the back projection technique [23,24] is applied, and the inverse STFT gives the separated signals.

In the proposed method, HPSS is utilized as a mask-generating function in the fourth line of Algorithm 1. Below, we explain this process, from input $\mathbf{z}$ to output $\mathcal{M}$, in a step-by-step manner.

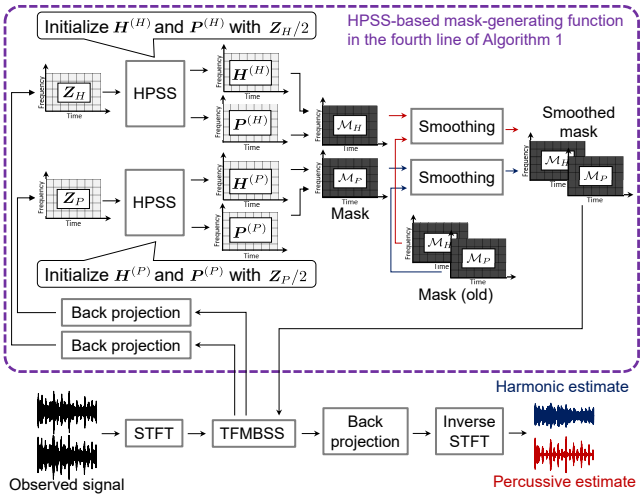

Fig. 1. Block diagram of the proposed method.

Firstly, the input auxiliary variable $\mathbf{z} \in \mathbb{C}^{2IJ}$ is split into two matrices, $\boldsymbol{Z}_H \in \mathbb{C}^{I\times J}$ and $\boldsymbol{Z}_P \in \mathbb{C}^{I\times J}$. Note that this splitting is merely reshaping of the variable, and no processing is applied. The half of $\mathbf{z}$ is always assigned to $\boldsymbol{Z}_H$, and the other half is assigned to $\boldsymbol{Z}_P$ for every iteration. Then, the back projection technique [23, 24] is applied to each of them in order to fix the frequency-wise scales. This scale alignment is crucial for HPSS to work properly.

Secondly, two independent HPSS processes are performed. Each HPSS is initialized by $\boldsymbol{Z}_H$ and $\boldsymbol{Z}_P$ as follows:

$$|\boldsymbol{B}^{(H)}|^{\xi} = |\boldsymbol{Z}_H|^{\xi}, \qquad |\boldsymbol{H}^{(H)}| = |\boldsymbol{P}^{(H)}| = |\boldsymbol{Z}_H|^{\xi}/2, \tag{11}$$

$$|\boldsymbol{B}^{(P)}|^{\xi} = |\boldsymbol{Z}_P|^{\xi}, \qquad |\boldsymbol{H}^{(P)}| = |\boldsymbol{P}^{(P)}| = |\boldsymbol{Z}_P|^{\xi}/2, \tag{12}$$

where the HPSS for $\boldsymbol{Z}_H$ is performed with $\boldsymbol{B}^{(H)}$, $\boldsymbol{H}^{(H)}$ and $\boldsymbol{P}^{(H)}$, and that for $\boldsymbol{Z}_P$ is performed with $\boldsymbol{B}^{(P)}$, $\boldsymbol{H}^{(P)}$ and $\boldsymbol{P}^{(P)}$. By iterating the update rules in (3)–(5), two pairs of separated signals are obtained: $|\boldsymbol{H}^{(H)}|^{2\rho}$ and $|\boldsymbol{P}^{(H)}|^{2\rho}$ are separated from $\boldsymbol{Z}_H$, and $|\boldsymbol{H}^{(P)}|^{2\rho}$ and $|\boldsymbol{P}^{(P)}|^{2\rho}$ are separated from $\boldsymbol{Z}_P$.

Thirdly, two Wiener-like masks [10], $\mathcal{M}_H$ and $\mathcal{M}_P$, are constructed using the results of HPSS as follows:

$$\mathcal{M}_H = \frac{|\boldsymbol{H}^{(H)}|^2}{|\boldsymbol{H}^{(H)}|^2 + |\boldsymbol{P}^{(H)}|^2}, \quad \mathcal{M}_P = \frac{|\boldsymbol{P}^{(P)}|^2}{|\boldsymbol{H}^{(P)}|^2 + |\boldsymbol{P}^{(P)}|^2}, \tag{13}$$

where all the operations are performed element-wise. These masks enhance the harmonic or percussive components by eliminating the other components. Therefore, the demixing filter is informed about which component to reduce.

Finally, these masks $\mathcal{M}_H \in [0,1]^{I\times J}$ and $\mathcal{M}_P \in [0,1]^{I\times J}$ are concatenated and vectorized to form a mask $\mathcal{M} \in [0,1]^{2IJ}$ that can be applied to $\mathbf{z} \in \mathbb{C}^{2IJ}$ as in the fifth line of Algorithm 1. These four steps define the function `generateMask`$(\cdot)$, which maps $\mathbf{z}$ to $\mathcal{M}$, for the proposed method. Note that the quality of masks depend on the degree of success of HPSS. Since HPSS may fail to correctly separate the signals, we additionally propose a stabilization technique that is separately explained in the next subsection.

3.3. Mask smoothing technique for TFMBSS

Since TFMBSS is built upon a proximal algorithm [10], its update should stay in the proximity of the previous state. However, a mask-generating function may violate this requirement, which makes the algorithm unstable. To stabilize the algorithm regardless of the choice of mask-generating function, we propose a mask smoothing technique for TFMBSS.

The algorithm can be stabilized by avoiding huge difference between the current and previous masks. Therefore, we propose the following smoothing rule that is independently applied to $\mathcal{M}_H$ and $\mathcal{M}_P$ right after the calculation of these masks in (13):

$$\mathcal{M} = \mathcal{M}^{\beta} \odot \mathcal{M}_{\mathrm{old}}{}^{\beta_{\mathrm{old}}} \tag{14}$$

where $\mathcal{M}_{\mathrm{old}}$ is the mask obtained in the previous iterate, and $\beta \geq 0$ and $\beta_{\mathrm{old}} \geq 0$ are the smoothing parameters such that $\beta + \beta_{\mathrm{old}} = 1$. The degree of smoothness is decided by this parameter. When $\beta_{\mathrm{old}} = 0$ (i.e., $\beta = 1$), (14) does nothing to the current mask. By increasing β_{old} toward 1, the smoothing effect becomes stronger. The effect of this smoothing will be discussed via an experiment.

Note that the applicability of this smoothing technique is not limited to the proposed method. It can be applied to any BSS method based on TFMBSS. Therefore, it should be beneficial for TFMBSS-based algorithms already proposed in the literature [9, 10].

Table 1. Experimental conditions

Window function in STFT	Hann window
Window length in STFT	128 ms
Shift length in STFT	64 ms
Parameters in HPSS	$\kappa_H = 1.02, \kappa_P = 1.01$ $\xi = 2, \rho = 1$
Parameters in TFMBSS	$\alpha = 0.25$ $\mu_1 = \mu_2 = 1.0$
Number of iterations in TFMBSS	500

4. EXPERIMENTS

4.1. Conditions

We conducted experiments of separating drums and other musical instruments. For the testing data, "drums" and "other" sources of 20 songs in the SiSEC 2016 MUS dataset [25] were used. To produce multichannel observed signals, these dry sources were convolved with the impulse response "E2A" in the RWCP database [26] (reverberation time was 300 ms) with 5.66 cm microphone spacing and source orientations of 50°&130° (90° corresponds to the normal direction of the two microphones), as in [27]. The improvement of source-to-distortion ratio (SDR) [28] was used as an evaluation score because it is in good agreement with both degree of separation and absence of artificial noise. The other experimental conditions are summarized in Table 1.

4.2. Effect of numbers of iterations in HPSS

The number of HPSS iterations (3)–(5) affects the separation performance of the proposed method. Here, we compare the performance of the proposed method with various numbers of iterations in HPSS. Table 2 shows the SDR improvements of the proposed method averaged over harmonic and percussive sources and 20 songs, where the smoothing parameters were set to $\beta = 0.25$ and $\beta_{\mathrm{old}} = 0.75$. Since the update rules (3)–(5) of HPSS rapidly converges to a local minimum of (1), the performance of the proposed method saturates when the number of iterations in HPSS is set to around 15. On the basis of this result, we set the number of iterations of HPSS for the proposed method to 15 in the following experiments.

Table 2. Average SDR improvements of the proposed method with various numbers of iterations in HPSS

Number of iterations in HPSS	Average SDR improvement [dB]
1	8.29
3	10.40
5	10.87
7	10.98
9	11.08
11	10.79
13	11.09
15	**11.29**
20	11.06

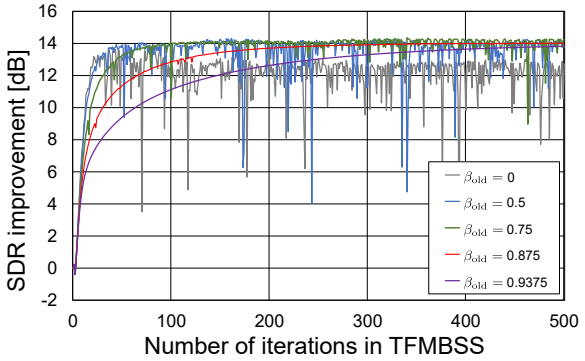

Fig. 2. Typical example of convergence behaviors in the proposed method with various β and $\beta_{\rm old}$.

Table 3. Average SDR improvements of the proposed method with various smoothing parameters β and $\beta_{\rm old}$

β	$\beta_{\rm old}$	Average SDR improvement [dB]
1	0	7.40
0.5	0.5	10.14
0.25	0.75	**11.29**
0.125	0.875	11.27
0.0625	0.9375	11.01

4.3. Effect of smoothing parameters

Here, we investigate the effect of the proposed mask-smoothing technique in Section 3.3. A typical example of the convergence behaviors of SDR improvements averaged over harmonic and percussive sources is shown in Fig. 2. Note that $\beta = 1 - \beta_{\rm old}$, and $\beta_{\rm old} = 0$ corresponds to the algorithm without the smoothing technique. From this figure, we can confirm that the smoothing process clearly stabilizes the proposed algorithm. This tendency was the same for the other songs.

As in the figure, there is a trade-off between stability and convergence speed. Therefore, for a fixed number of iterations, the smoothing parameter should affect the separation performance. Table 3 summarizes the SDR improvements averaged over 20 songs. This table shows that excess amount of smoothing results in degradation of performance. A preferable condition of the smoothing process seems around $\beta = 0.25$ or 0.125 for the proposed method. The comparison in the next subsection is made by $\beta = 0.25$.

4.4. Comparison with existing HPSS and BSS algorithms

We compared five methods: (a) single-channel HPSS [12, 13], (b) multichannel HPSS [21], (c) IVA based on the auxiliary function technique (AuxIVA) [6], (d) ILRMA [7, 8], and (e) the proposed method. For the multichannel HPSS, we used the MATLAB implementation[1] provided by the authors of [21], where two conditions of parameters (given and fine-tuned) were used as shown in Table 4 (see [21] for details of these parameters). For single-channel HPSS, AuxIVA, ILRMA, and the proposed method, we used 128-ms-long Hann window with half shifting. The other conditions for the conventional methods are summarized in Table 4.

Table 4. Conditions for other HPSS and BSS algorithms

Parameters for multichannel HPSS described in [21]	128 ms Hann window with 1/2 shift $\alpha_h = \alpha_p = 10$, $m_h = m_p = 5$ $\gamma_1 = 0.5$, $\gamma_2 = 1$
Fine-tuned parameters for multichannel HPSS	16 ms Hann window with 1/2 shift $\alpha_h = \alpha_p = 5$, $m_h = m_p = 10$ $\gamma_1 = \gamma_2 = 1$
Number of bases in ILRMA	10 for each source
Number of iterations	20 for single-channel HPSS 15 for multichannel HPSS 30 for AuxIVA 100 for ILRMA

Table 5. Average SDR improvement for HPSS and BSS algorithms

HPSS/BSS algorithm	Average SDR improvement [dB]
Single-channel HPSS	4.80
Multichannel HPSS with the parameters in [21]	-0.47
Multichannel HPSS with the fine-tuned parameters	1.25
AuxIVA	7.91
ILRMA	7.76
Proposed method	**11.29**

Table 5 shows the average SDR improvements for all the methods. In this experiment, the single-channel HPSS outperformed the multichannel one. While multichannel HPSS can reduce the artificial distortions in the estimated signals using spatial information, the degree of separation tends to be sacrificed. Also, multichannel HPSS aims to separate the typical stereo music with panning mixtures. Although the spatial covariance model [29] in multichannel HPSS can handle the convolutive mixture (9), its parameter optimization was unstable and the performance was not high in this experiment. AuxIVA and ILRMA provide better performance compared to the conventional HPSS methods. This is because the source models assumed in IVA and ILRMA work well for modeling the drums and instrumental signals. The proposed method outperformed the other HPSS and BSS techniques because the suitable source models for separating the harmonic and percussive sources are utilized.

5. CONCLUSION

In this paper, a new algorithm that unifies TFMBSS and HPSS-based time-frequency mask estimation was proposed. Also, we revealed the importance of the mask-smoothing process to greatly improve the performance of TFMBSS. The proposed method outperformed the conventional HPSS-based and FDICA-based techniques.

The utilization of another time-frequency mask is our future work. We expect that the proposed mask-smoothing technique is useful even in such applications with various time-frequency masks.

[1] Retrieved from https://www.irisa.fr/metiss/ngoc/sw/hpss.rar

6. REFERENCES

[1] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomput.*, vol. 22, pp. 21–34, 1998.

[2] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Convolutive blind source separation for more than two sources in the frequency domain," *Proc. ICASSP*, 2004, pp. III-885–III-888.

[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. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

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

[5] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

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

[7] 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. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[8] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," *In Audio Source Separation,* S. Makino, Ed., pp. 125–155, Springer, Cham, 2018.

[9] K. Yatabe and D. Kitamura, "Time-frequency-masking-based determined BSS with application to sparse IVA," *Proc. ICASSP*, 2019, pp. 715–719.

[10] K. Yatabe and D. Kitamura, "Determined BSS based on time-frequency masking and its application to harmonic vector analysis," *arXiv:2004.14091*, 2020.

[11] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," *Proc. EUSIPCO*, 2008.

[12] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," *Proc. ICASSP*, 2010, pp. 425–428.

[13] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Comparative evaluations of various harmonic/percussive sound separation algorithms based on anisotropic continuity of spectrogram," *Proc. ICASSP*, 2012, pp. 465–468.

[14] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Trans. ASLP*, vol. 22, no. 1, pp. 228–237, 2012.

[15] H. Tachibana, N. Ono, H. Kameoka, and S. Sagayama, "Harmonic/percussive sound separation based on anisotropic smoothness of spectrograms," *IEEE/ACM Trans. ASLP*, vol. 22, no. 12, pp. 2059–2073, 2014.

[16] K. Yatabe and D. Kitamura, "Determined blind source separation via proximal splitting algorithm," *Proc. ICASSP*, 2018, pp. 776–780.

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

[18] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[19] 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, 2015.

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

[21] N. Q. K. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity," *Proc. ICASSP*, 2011, pp. 205–208.

[22] 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. SAP*, vol. 12, no. 5, pp. 530–538, 2004.

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

[24] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *Proc. ICA*, 2001, pp. 722–727.

[25] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," *Proc. LCA/ICA*, 2017, pp. 323–332.

[26] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. LREC*, 2000, pp. 965–968.

[27] D. Kitamura, N. Ono, and H. Saruwatari, "Experimental analysis of optimal window length for independent low-rank matrix analysis," *Proc. EUSIPCO*, 2017, pp. 1210–1214.

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

[29] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.