BLIND SPARSE SOURCE SEPARATION WITH SPATIALLY SMOOTHED TIME-FREQUENCY MASKING

[1,2]*Shoko Araki,* [1]*Hiroshi Sawada,* [1]*Ryo Mukai, and* [1,2]*Shoji Makino*

shoko@cslab.kecl.ntt.co.jp
[1]NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
[2]Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo-shi, Hokkaido 060-0814, Japan

ABSTRACT

In this paper, we handle the blind sparse source separation problem with time-frequency masks. Some existing methods employ time-frequency *binary* masks to extract the signals, and therefore, the extracted signals tend to contain loud musical noise. It causes from the winner-take-all property of the binary mask. To overcome the problem, this paper proposes some non-binary time-frequency masks that allow each time-frequency component to belong to several output signals. Experimental results show that our proposed method can separate signals with little distortion.

1. INTRODUCTION

A time-frequency mask approach to the blind sparse source separation (BSS) is recently widely studied (e.g., [1, 2]). The time-frequency mask approach is attractive because it can handle an underdetermined problem where the sources outnumber the sensors. With regards to these problems, we also have already proposed a time-frequency mask approach that is based on the observation vector clustering [3].

The time-frequency mask methods rely on the assumption of source sparseness. If signals are sufficiently sparse, that is, most of the samples of each signal are almost zero, we can assume that the sources rarely overlap. The time-frequency mask approaches employ this assumption and they usually extract each signal by applying a time-frequency *binary* mask to the observed mixture. By using the time-frequency binary mask we can extract each signal from a mixture. However, the use of binary masks leads to too much discontinuous zero-padding to the extracted signals, and therefore, they tend to contain loud musical noise, which is undesirable for audio applications. The zero-padding is caused by the binary mask's *winner-take-all* property: each time-frequency point of the observed mixture is allowed to belong to only one extracted signal.

In order to reduce the musical noise problem, in this paper, we propose some non-binary time-frequency masks that allow each time-frequency component to belong to several output signals. Our non-binary masks are designed by using the cluster information of the observation vector clustering [3], therefore, the mask has the spatially-smoothness. We show that our proposed non-binary masks can reduce musical noise with no degradation of separation performance.

2. PROBLEM DESCRIPTION

Suppose that sources $\mathrm{s}_1, \ldots, \mathrm{s}_N$ are convolutively mixed and observed at M sensors

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_l \mathrm{h}_{jk}(l)\, \mathrm{s}_k(t-l),\ j=1,\ldots,M, \quad (1)$$

where $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to sensor j. In this paper, we look especially at a situation where the number of sources N can exceed the number of sensors M ($N > M$). We assume that N and M are known, and that the sensor alignment does not cause the spatial aliasing problem. Our goal is to obtain separated signals $\mathrm{y}_k(t)$ that are estimations of s_k calculated solely from M observations.

This paper employs a time-frequency domain approach. Using a short-time Fourier transform (STFT), the convolutive mixtures (1) can be converted to instantaneous mixtures at each frequency f:

$$x_j(f,\tau) \approx \sum_{k=1}^{N} h_{jk}(f) s_k(f,\tau), \quad (2)$$

or in vector notation,

$$\mathbf{x}(f,\tau) \approx \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,\tau), \quad (3)$$

where $h_{jk}(f)$ is the frequency response from source k to sensor j, $s_k(f,\tau)$ is the STFT of a source signal s_k, and τ is a time index. We call $\mathbf{x} = [x_1, \ldots, x_M]^T$ *an observation vector* and $\mathbf{h}_k = [h_{1k}, \ldots, h_{Mk}]^T$ is a vector of the frequency responses from source s_k to all sensors.

We assume the sparseness of sources in the time-frequency domain. This assumption has been widely employed for solving the underdetermined problem (e.g. [1, 2, 3]). When the signals are sufficiently sparse, we can assume that the sources rarely overlap at each time-frequency point, and (3) can be approximated as

$$\mathbf{x}(f,\tau) \quad \approx \quad \mathbf{h}_k(f) s_k(f,\tau), \quad k \in \{1, \cdots, N\}, \quad (4)$$

where $s_k(f,\tau)$ is a dominant source at the time-frequency point (f,τ). For instance this is true for speech signals in the time-frequency domain [1].

3. CONVENTIONAL BINARY MASK APPROACH

3.1. Method

Here we employ the method proposed in [3]. First, we normalize all observation vectors $\mathbf{x}(f,\tau)$ for all time-frequency points

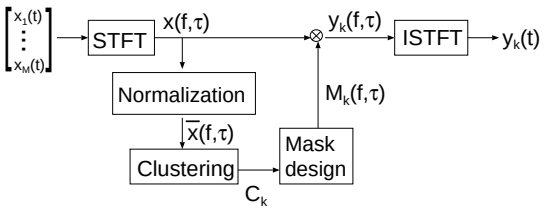

Figure 1: Flow of the time-frequency mask approach.

(f,τ). The normalization includes phase-normalization with respect to a sensor J and the frequency-normalization,

$$\bar{x}_j(f,\tau) \leftarrow |x_j(f,\tau)| \exp\left[\jmath \frac{\arg[x_j(f,\tau)/x_J(f,\tau)]}{4fc^{-1}d_{\max}}\right] \quad (5)$$

where c is the propagation velocity and $d_{\max}$ is the maximum distance between sensor J and a sensor ${}^{\forall}j \in \{1,\ldots,M\}$ [3]. Then, we apply unit-norm normalization

$$\bar{\mathbf{x}}(f,\tau) \leftarrow \bar{\mathbf{x}}(f,\tau) \,/\, ||\bar{\mathbf{x}}(f,\tau)|| \quad (6)$$

to $\bar{\mathbf{x}}(f,\tau) = [\bar{x}_1(f,\tau),\ldots,\bar{x}_M(f,\tau)]^T$.

Then we find clusters $C_1,\ldots,C_N$ formed by all normalized vectors $\bar{\mathbf{x}}(f,\tau)$. The clustering criterion is to minimize the total sum $\mathcal{J}$ of the squared distances d_k^2 between cluster members and their centroid:

$$\mathcal{J} = \sum_{k=1}^{M} \sum_{\bar{\mathbf{x}}(f,\tau)\in C_k} d_k^2(f,\tau) \quad (7)$$

$$d_k(f,\tau) = ||\bar{\mathbf{x}}(f,\tau) - \mathbf{c}_k||. \quad (8)$$

The clustering is realized by the following iterative updates:

$$C_k = \{\bar{\mathbf{x}}(f,\tau) \mid k = \mathrm{argmin}_i d_i^2(f,\tau)\} \quad (9)$$

$$\mathbf{c}_k \leftarrow E[\bar{\mathbf{x}}(f,\tau)]_{\bar{\mathbf{x}}\in C_k}, \quad \mathbf{c}_k \leftarrow \mathbf{c}_k/||\mathbf{c}_k||, \quad (10)$$

where $E[\cdot]_{\bar{\mathbf{x}}\in C_k}$ is a mean operator for the members of a cluster C_k. This minimization can be performed efficiently with the k-means clustering algorithm [4] with a given source number N.

Because each resulting cluster corresponds to an individual source, finally, we have separated signals $y_k(f,\tau) = M_k(f,\tau)x_j(f,\tau)$ where

$$M_k(f,\tau) = \begin{cases} 1 & \bar{\mathbf{x}}(f,\tau) \in C_k \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

then we have time-domain outputs $\mathrm{y}_k(t)$ by using an inverse STFT (ISTFT).

3.2. Rationale

Let assume that an impulse response $h_{jk}(f)$ is modeled as

$$h_{jk}(f) \approx \lambda_{jk} \exp[-\jmath 2\pi f \tau_{jk}]. \quad (12)$$

where $\lambda_{jk} \geq 0$ and τ_{jk} are the attenuation and the time delay from source k to sensor j. If we assume an anechoic situation and no-spatial aliasing, the parameters λ_{jk} and τ_{jk} are determined solely by the geometric condition of the sources and sensors. Considering (4) and (12), the normalized vector $\bar{\mathbf{x}}(f,\tau)$ can be written as

$$\bar{x}_j(f,\tau) \approx \frac{\lambda_{jk}}{A_k} \exp\left[-\jmath \frac{\pi(\tau_{jk}-\tau_{Jk})}{2c^{-1}d_{\max}}\right], \quad (13)$$

where $A_k = \sqrt{\sum_{j=1}^{M}\lambda_{jk}^2}$. We can see that the normalized observation vector $\bar{\mathbf{x}}(f,\tau)$ depends only on the source geometry λ_{jk} and τ_{jk} of the source s_k, which is dominant at the time-frequency point (f,τ). Therefore, the normalized observation vectors can be clustered based on the source geometry. In other words, the observation vectors $\bar{\mathbf{x}}(f,\tau)$ in a cluster C_k are assumed to be the signal components coming from a physically spatially localized region.

4. NON-BINARY MASK PROPOSALS

The binary mask (11) can extract each signal from a mixture, however, it causes much discontinuity in the extracted signals, and therefore, the musical noise occurs. It arises as a result of the binary mask's *winner-take-all* property, as mentioned in Section 1. In order to relax the property, in this section, we propose some *non-binary* time-frequency masks that allow each time-frequency component to belong to several output signals.

Here, we specify the mask based on the distances $d_k(f,\tau)$ (8) between cluster members and their centroid, which was utilized in the clustering procedure (9). From the fact mentioned in Sec. 3.2, our proposing non-binary masks become a *spatially smooth* mask.

4.1. MASK1: Sigmoid based mask

The first mask has the shape of the sigmoid function [5]. The mask is defined as

$$M_k(f,\tau) = M_k(d_k(f,\tau)) = \frac{1}{1+\exp\left[g(d_k(f,\tau)-th_k)\right]} \quad (14)$$

where th_k and g are parameters deciding the shape of the sigmoid function. The smaller th_k, the more interference power is suppressed, but the more musical noise sound in the extracted signals. An example sigmoid mask is shown in Fig. 2.

4.2. MASK2: Baysian theorem based mask

As the second mask, we newly propose to use the Bayse's theorem by assuming each cluster member $\bar{\mathbf{x}}(f,\tau) \in C_k$ has a complex Gaussian distribution,

$$p(\bar{\mathbf{x}}|C_k) = \frac{1}{(2\pi)^N|\mathbf{\Sigma}_k|} \exp\left[-(\bar{\mathbf{x}}-\mathbf{c}_k)^H \mathbf{\Sigma}_k^{-1}(\bar{\mathbf{x}}-\mathbf{c}_k)\right] \quad (15)$$

where $\mathbf{\Sigma}_k$ is a covariance matrix of $\bar{\mathbf{x}}(f,\tau) \in C_k$. Note that the term $(\bar{\mathbf{x}}-\mathbf{c}_k)^H\mathbf{\Sigma}_k^{-1}(\bar{\mathbf{x}}-\mathbf{c}_k)$ in (15) is the square distance d_k^2 normalized by $\mathbf{\Sigma}_k$. With the Bayse's theorem, the posterior probability of $\bar{\mathbf{x}}(f,\tau)$ being $\bar{\mathbf{x}}(f,\tau) \in C_k$ is given as

$$P\left(C_k|\bar{\mathbf{x}}(f,\tau)\right) = \frac{p\left(\bar{\mathbf{x}}(f,\tau)|C_k\right)P(C_k)}{p(\bar{\mathbf{x}}(f,\tau))} \quad (16)$$

where $P(C_k)$ is known priori (here we assume $P(C_k) = 1/N, {}^{\forall}k$), and $p(\bar{\mathbf{x}}(f,\tau)) = \sum_{k=1}^{N} p(\bar{\mathbf{x}}(f,\tau)|C_k)P(C_k)$.

Here we use this posterior as our mask,

$$M_k(f,\tau) = P\left(C_k|\bar{\mathbf{x}}(f,\tau)\right) \quad (17)$$

The mask has a property that $\sum_{k=1}^{N} M_k(f,\tau) = 1$, because of the normalization with the denominator of (16). An example

mask is shown in Fig. 3. It looks scattered. This is because even if $d_k(f,\tau) = d_k(f',\tau')$, it does not mean $\bar{\mathbf{x}}(f,\tau) = \bar{\mathbf{x}}(f',\tau')$ and therefore $p(\bar{\mathbf{x}}(f,\tau)|C_k) \neq p(\bar{\mathbf{x}}(f',\tau')|C_k)$.

4.3. MASK3: Directivity pattern based mask

If we know the sensor configuration, we can employ the directivity pattern of a fixed null-beamformer (NBF) that makes nulls towards the interference directions [6] as our non-binary mask. This mask realizes a spatially smooth mask, straightforwardly.

STEP1: Direction estimation
First, we have to estimate the direction of arrivals (DOAs) $\mathbf{q}_k$ of all N sources, where $\mathbf{q}_k$ is a 3-dimensional vector of a unit-norm representing the direction of the source s_k. DOAs are estimated by the method described in [7]

$$\hat{\mathbf{q}}_k = -\frac{2d_{\max}}{\pi}\boldsymbol{P}^{+}\arg\left[\mathbf{c}_k\right] \tag{18}$$

where $\boldsymbol{P} = [\mathbf{p}_1 - \mathbf{p}_J, \cdots, \mathbf{p}_M - \mathbf{p}_J]^T$ and $\boldsymbol{p}_j$ is a 3-dimensional vector representing the location of sensor j.

STEP2: Mask design by the directivity pattern of NBF
Then, we make an NBF that makes nulls towards $N-1$ interference directions. Remember that we have only $M < N$ sensors in an underdetermined case. In order to form $N-1$ nulls, we assume $V > (N-1)+1$ (virtual) sensors [6] of arbitral positions of V virtual sensors $\mathbf{p}'_j$ $(j = 1, \cdots, V)$.

Then NBF $\boldsymbol{W}(f)$ is calculated by $\boldsymbol{W}(f) = \boldsymbol{H}^{-1}_{\text{NBF}}(f)$, where $\boldsymbol{H}_{\text{NBF}}(f)$ is a $(V \times V)$ matrix whose ji-th element $H_{\text{NBF}ji}(f) = \exp\left[-\jmath 2\pi f c^{-1}(\mathbf{p}'_j - \mathbf{p}'_J)^T\hat{\mathbf{q}}_i\right]$, $\hat{\mathbf{q}}_1 = \hat{\mathbf{q}}_k$, and k is the index of the source to be extracted. The directivity pattern of the NBF is obtained as a function of DOA $\mathbf{q}$

$$F_k(f,\mathbf{q}) = \sum_{j=1}^{V} W_{1j}(f)\exp\left[-\jmath 2\pi f c^{-1}(\mathbf{p}'_j - \mathbf{p}'_J)^T\mathbf{q}\right]. \tag{19}$$

An example gain pattern of (19) is given by Fig. 4 (a).

By employing this gain pattern, our proposed mask becomes

$$M_k(f,\tau) = F_k(f, \mathbf{q}(f,\tau)) \tag{20}$$

where $\mathbf{q}(f,\tau)$ is the estimated DOA of a observation $\bar{\mathbf{x}}(f,\tau)$ at each time-frequency [7].

We can modify the mask so that it has a small constant gain except for a main beam [6]. Figure 4 (b) shows an example.

5. EXPERIMENTS

5.1. Experimental conditions

We evaluated the performance with 5-second English speech sources and measured impulse responses in a room (Fig. 5). The observations were simulated by (1). The sampling rate was 8 kHz, the STFT frame size $L = 512$ and the frame shift was $L/2$, $L/4$ and $L/8$. Here, we only tested the 3-microphones and 4-sources case.

The separation performance was evaluated in terms of the signal-to-interference ratio (SIR) improvement [3]. Moreover, we also evaluated the separated sound quality with the signal to distortion ratio (SDR) [3]. We investigated four speaker combinations and averaged the results for all outputs. Furthermore, we

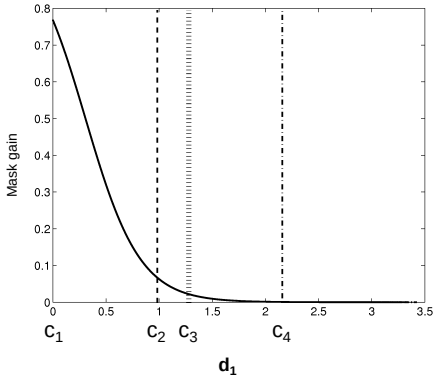

Figure 2: Sigmoid mask (MASK1). The horizontal axis is the distance d_1 from the centroid $\boldsymbol{c}_1$.

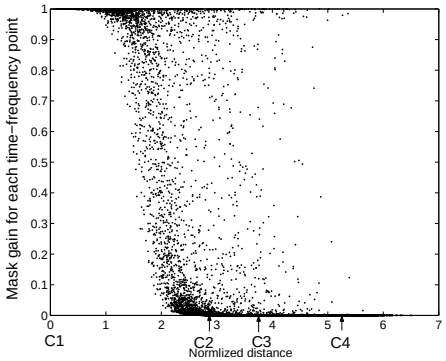

Figure 3: Baysian mask (MASK2). The horizontal axis is the normalized distance from the centroid $\boldsymbol{c}_1$: $\sqrt{(\bar{\mathbf{x}} - \mathbf{c}_1)^H\boldsymbol{\Sigma}_1^{-1}(\bar{\mathbf{x}} - \mathbf{c}_1)}$, see section 4.2.

conducted a small subjective test with four listeners, since SDR cannot evaluate musical noise [8]. It was a paired comparison test, where we paired up a conventional binary mask output with a proposed smooth mask output, and each listener judged which has less musical noise for many pairs.

5.2. MASKs' parameters

For MASK1, g was 7.8, that realized as good SIR as with a conventional binary mask, and th_k was the standard deviation of $d_k(f,\tau) \in C_k$. For MASK2, $P(C_k) = 1/N, {}^{\forall} k$, that is we assume the same priori for all k. For MASK3, the number of the virtual sensors $V = 4$.

5.3. Results

Results are shown in Table 1, where "BM" denotes the conventional binary mask (11), and MASK3' means the modified MASK3 with the same way as Fig. 4(b). By using non-binary masks, we improved the SDR without degrading the SIR. Table 1 also have the result of the listening test with paired comparison (PC) for the frame shift=$L/4$. The PC values show the percentage of MASK's superiority over the BM in musical noise

$$\text{PC} = 100 \times \frac{\text{\# of pairs where MASK is better than BM}}{\text{\# of all pairs in the test}}$$

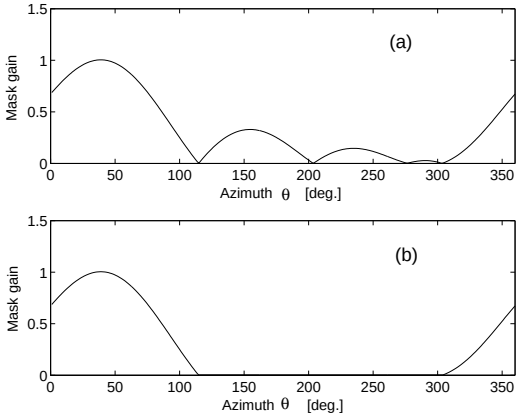

Figure 4: (a) Example directivity based mask (MASK3) and (b) its modification (MASK3') for extracting s_1 in a setting shown in Fig. 5. The horizontal axis is the azimuth θ whose definition is shown in Fig. 5.

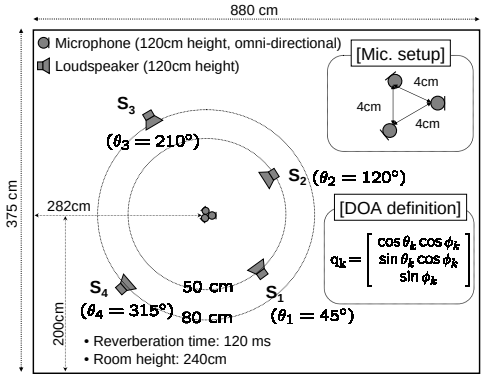

Figure 5: Experimental setup. θ_k ($k = 1, \cdots, 4$) are the azimuth values of signals. The elevation $\phi_k = 0, \forall k$.

where the numerator means # of pairs where a MASK{1,2,3,3'} output has less musical noise than a conventional binary mask output. The listening test results show that our non-binary masks decrease the musical noise significantly.

The observed phenomena were as follows.

1. MASK1 achieved good performance in spite of its simple realization (14). It has only one parameter g and its setting is not difficult.
2. The ability to reduce musical noise of MASK2 (baysian mask) was smaller than these of other masks. This is because that the MASK2 is a non-binary mask but it does not have spatial continuity shown in Fig. 3.
3. SIR of MASK3 was not good because MASK3 has large gains between null directions (see Fig. 4(a)). On the other hand, MASK3' (Fig. 4(b)) realized good performance in SIR with less musical noise.
4. With our non-binary masks the interferences in outputs were now understandable, although the effect is not reflected in SIR. This was because the interferences' residual also stayed without zero-padding.
5. Our non-binary masks worked effectively when the frame shift was a half or quarter of a frame size (Shift = $L/2$ or $L/4$). For the fine frame shift (Shift=$L/8$), the musical noise of BM was not so loud, and therefore, our non-binary masks had small impact.

Table 1: Experimental results. SIR and SDR are in dB. "Shift" stands for the frame shift. "PC" is the result of the listening test when Shift=$L/4$.

		Shift$L/2$	Shift$L/4$	Shift$L/8$	PC(%)
BM	SIR_i	13.5	14.5	14.9	-
	SDR_i	8.1	8.9	9.1	
MASK1	SIR_i	13.4	14.6	14.7	90.6
	SDR_i	8.4	8.6	8.7	
MASK2	SIR_i	14.2	15.1	15.3	62.5
	SDR_i	8.5	9.2	9.3	
MASK3	SIR_i	12.1	13.0	13.2	87.5
	SDR_i	8.5	9.3	9.4	
MASK3'	SIR_i	13.1	13.9	14.2	81.3
	SDR_i	8.9	9.6	9.8	

6. CONCLUSION

We proposed spatially smooth non-binary masks which allow each time-frequency component to belong to several output signals. Although they do not assure the perfect time-continuity of the output signals, the proposed masks effectively reduced the musical noise without decreasing the separation performance.

7. REFERENCES

[1] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[2] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[3] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. IWAENC 2005*, Sept. 2005.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley Interscience, 2nd edition, 2000.

[5] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of a dominant source signal from mixtures of many sources," in *Proc. ICASSP2005*, Mar. 2005, vol. III, pp. 61–64.

[6] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Source extraction from speech mixtures with null-directivity pattern based mask," in *Proc. HSCMA2005*, Mar 2005, pp. d1–d2.

[7] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. ICASSP2006*, May 2006, vol. 5, pp. 33–36.

[8] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. on SAP*, vol. 7, no. 2, pp. 126–137, 1999.