# SPARSE VECTOR FACTORIZATION FOR UNDERDETERMINED BSS USING WRAPPED-PHASE GMM AND SOURCE LOG-SPECTRAL PRIOR

*Shoko Araki* *Tomohiro Nakatani*

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

**ABSTRACT**

We propose a sparse vector factorization (SVF) approach for blind source separation, which inherently avoids the permutation problem. The SVF assumes the sparseness of sources, and defines a sparse vector (SV) that consists of the locational and spectral features of each source at all the frequencies. Then, by assuming that the locational and spectral SVs are generated by frequency-independent parameters, the method executes the SVF. Our locational feature is the phase difference (PD) between two microphone observations, and we model it with a frequency-independent time-difference of arrival (TDOA) parameter. Moreover, we employ the wrapped-phase GMM in order to take the spatial aliasing problem into account. On the other hand, the spectral feature is the log spectrum, and we provide a prior for a spectral parameter. The SVF is formulated with a maximum a posteriori (MAP) estimation framework, where the locational and spectral parameters are inferred by the EM algorithm. Experimental results show that our proposed method can separate signals successfully even for an underdetermined case.

***Index Terms***— Source separation, sparse sources, vector factorization, log spectrum, EM algorithm

## 1. INTRODUCTION

Blind source separation (BSS) is an approach for estimating source signals based only on the mixed signal information observed at each microphone. For the underdetermined BSS, time-frequency (t-f) mask based approaches that cluster the t-f components according to the sparseness of sources have been widely studied (e.g., [1–3]). Most of the t-f mask approaches employ a two-stage approach: the first stage relies on the source locational features and separates the signals by clustering the t-f components at each frequency bin, and then the second stage solves the permutation problem among frequencies, utilizing the source locational or spectral information. The permutation solution in the second stage usually requires an additional effort to obtain high separation performance.

This paper proposes an approach that inherently prevents the permutation problem. We refer to our method as the sparse vector factorization (SVF). The sparse vector (SV) consists of the locational and spectral features of each source at all the frequencies, which are generated by *frequency-independent parameters* (see Fig. 1). The SVF is an extention of the t-f component clustering, by which a pair of frame-wise observation vectors are factorized into SVs of individual sources based on the frequency independent parameters, and thus it does not cause the permutation problem. Moreover, this approach realizes high-performance BSS because the method relies on both the locational and spectral features.

The SVF is not a completely new idea, and indeed the same kind of ideas have already appeared in existing methods [4, 5]. The method in [4], utilized only the locational SVs of sources. To handle all the frequency components simultaneously, we modeled the phase difference (PD) between two microphone observations with a frequency-independent time-difference of arrival (TDOA) parameter. A phase wrapping problem of the PD in case of spatial aliasing was also considered by exploiting the wrapped-phase GMM [6]. Thanks to this modeling, the method [4] can cluster the t-f components without encountering the permutation problem even for an underdetermined case where the spatial aliasing occurs. However, since this method relies only on the locational information, performance at the frequencies where the PD of two sources lap over each other was insufficient.

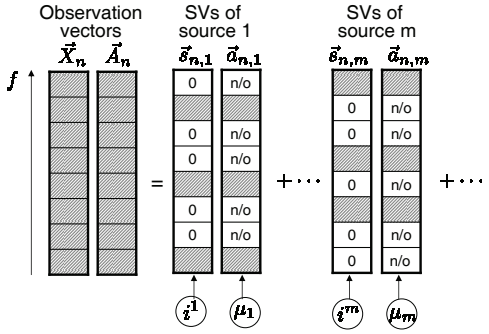


**Fig. 1**. Concept of SVF at a time frame $n$, where $\vec{s}_{n,m}$ and $\vec{a}_{n,m}$ are spectral and locational sparse vectors (SVs), respectively, and $i^m$ and $\mu_m$ are frequency-independent parameters for a source $m$. "n/o" means "not observed".

In the other method [5] based on the SVF concept, we showed how we can simultaneously cluster both the locational and spectral SVs. With this method, the variance of the spectral SV was modeled with a common amplitude modulation (AM) parameter which is shared by all the frequency components. However, because this spectral modeling was too naive to describe the fine structure of sources, e.g., speech signals, it was difficult to apply this method to an underdetermined case.

This paper proposes an SVF method that overcomes the above-mentioned shortcomings of [4, 5]. With the proposed method, the locational model follows that in [4], and the spectral model employs a more elaborate one. In concrete terms, the prior of the spectral SV is provided by a log spectral model that is capable of describing the fine structure of the speech signals (e.g., [7, 8]). The log spectral prior is given by an offline training. The SVF will be formulated with a maximum a posteriori (MAP) estimation framework, where the locational and spectral parameters are inferred by the EM algorithm. The experimental results show that our proposed method can separate speech signals even for an underdetermined case, and even at the frequencies where the PDs of two sources overlap due to the spatial aliasing.

## 2. PROBLEM DESCRIPTION

Suppose that $n(=1,...,N_n)$ and $f(=1,...,N_f)$ are time and frequency indices of a t-f slot, and that $N_s \geq 2$ speech signals $s_{n,f,1}, \cdots, s_{n,f,N_s}$ are mixed and observed at $N_m$ microphones,

$$x_{n,f,l} = \sum_{m=1}^{N_s} h_{f,l,m} s_{n,f,m}, \quad (1)$$

where $h_{f,l,m}$ is the frequency response from source $m$ to microphone $l$, $s_{n,f,m}$ is the STFT of a source $\mathrm{s}_m(t)$, $f \in \{0, \frac{1}{N_f} f_s, \ldots, \frac{N_f-1}{N_f} f_s\}$ is a frequency ($f_s$ is the sampling frequency) and $n \in \{0, \cdots, N_n - 1\}$ is a time-frame index. In this paper, without loss of generality, we handle a stereo case $N_m = 2$.

The separated signals $y_{n,f,m}$ are obtained with t-f masks $M_{n,f,m}$, which extract t-f points of the $m$-th source:

$$y_{n,f,m} = M_{n,f,m} x_{n,f,1}. \quad (2)$$

To estimate the t-f masks $M_{n,f,m}$, this paper assumes the sparseness of the sources [1], that is, at most one source is dominant at each t-f slot. With this assumption, we rewrite (1) as follows:

$$s_{n,f,m} = \begin{cases} x_{n,f,1} & \text{for } m = l_{n,f} \\ 0 & \text{for } m \neq l_{n,f} \end{cases} \quad (3)$$

where $l_{n,f}$ is the dominant source index at the t-f slot $(n, f)$. In (3), we assume that $|h_{f,1,m}| = 1$ and $\arg(h_{f,1,m}) = 0$, without loss of generality.

## 3. SPARSE VECTOR FACTORIZATION (SVF)

In this paper, the locational feature is the PDs between two channels $A_{n,f} = \arg[\frac{x_{n,f,2}}{x_{n,f,1}}]$, and the spectral feature is the complex spectrum $X_{n,f} = x_{n,f,1}$. We also define the observation vectors $\vec{A}_n$ and $\vec{X}_n$ (see Fig. 1), where a vector $\vec{\cdot}_n$ consists of $\cdot_{n,f}$ at all the frequencies at frame $n$.

Letting $\theta_A$ and $\theta_X$ be a set of locational and spectral model parameters (detailed in the following subsections), respectively, and $\{\vec{\cdot}_n\}$ be a set of vectors $\vec{\cdot}_n$ for all the time frames, we model the observation vectors $\vec{X}_n$ and $\vec{A}_n$ with a mixture model:

$$p(\{\vec{X}_n\}, \{\vec{A}_n\}, \theta_X, \theta_A) = \prod_{n,f} \sum_{m}^{M} p(X_{n,f}, A_{n,f}|m, \theta_X, \theta_A) p(\theta_X) p(\theta_A) p(m) \quad (4)$$

where we denote "$l_{n,f} = m$" by "$m$", and $p(m)$ is the mixture weight ($\sum_m p(m) = 1$). Assuming that $X_{n,f}$ and $A_{n,f}$ are mutually independent given $m$, we write $p(X_{n,f}, A_{n,f}|m, \theta_X, \theta_A)$ as

$$p(X_{n,f}, A_{n,f}|m, \theta_X, \theta_A) = p(X_{n,f}|m, \theta_X) \cdot p(A_{n,f}|m, \theta_A) \quad (5)$$

where $p(X_{n,f}|m, \theta_X)$ and $p(A_{n,f}|m, \theta_A)$ are the spectral (Sec. 3.1) and the locational model (Sec. 3.2), respectively. Moreover, $p(\theta_X)$ and $p(\theta_A)$ are the prior for the locational and spectral model parameters, respectively. 1

### 3.1. Spectral model

#### *3.1.1. Source and observation models*

This section defines the spectral model $p(X_{n,f}|m, \theta_X)$. From the sparseness assumption (3), the complex spectrum of each source $\vec{s}_{n,m}$ becomes a sparse vector. That is, $\vec{s}_{n,m}$ is the spectral SV (see Fig. 1). We model $s_{n,f,m}$ with a complex Gaussian distribution $N_c$:

$$p(s_{n,f,m}) = N_c(s_{n,f,m}; 0, \gamma^2_{n,f,m}) \quad (6)$$

where $\gamma^2_{n,f,m}$ is the variance of the source spectrum $E[|s_{n,f,m}|^2]$ at each t-f point. Moreover, we interpret the sparseness assumption in (3) as follows: two types of observations, $x_{n,f,1}$ and zeros, are always obtained simultaneously in relation to the dominant source and the other non-dominant sources. This interpretation can be represented by

$$p(X_{n,f}|m, \theta_X) = p(s_{n,f,m} = x_{n,f,m}) \prod_{m' \neq m}^{M} p(s_{n,f,m'} = 0) \quad (7)$$

The derivation of this model and its generative model can be found in [5].

#### *3.1.2. Spectral variance prior with a GMM*

In our previous paper [5], we model the spectral variance $\gamma^2_{n,f,m}$ by considering a common AM structure across frequencies:

$$\gamma_{n,f,m} = |X_{n,f}| \gamma'_{n,m}, \quad (8)$$

where $\gamma'_{n,m}$ is the time-variant spectral envelope which models the synchrony across frequencies, and $|X_{n,f}|$ represents the spectral fine structure. This model was really naive, and therefore, it was difficult to apply this model to the case of $N_s \geq 2$.

Instead of using this naive model, this paper proposes to utilize a more elaborate model with a log spectral model. First, we introduce a log spectral parameter

$$\rho_{n,f,m} = \log(\gamma^2_{n,f,m}), \quad (9)$$

and consider a vector $\vec{\rho}_{n,m}$ in order to model the source spectrum. Moreover, we provide a prior $p(\theta_X) = p(\vec{\rho}_{n,m})$ for this log spectral parameter at each source $m$ with a mixture of $I$ Gaussians:

$$\begin{aligned} p(\vec{\rho}_{n,m}) &= \sum_{i^m}^{I} p(\vec{\rho}_{n,m}|i^m) p(i^m) \\ &= \sum_{i^m}^{I} p(i^m) \prod_f N(\rho_{n,f,m}; \nu_{f,i^m}, v_{f,i^m}) \end{aligned} \quad (10)$$

where the mean $\nu_{f,i^m}$, the variance $v_{f,i^m}$, and the weight $p(i^m)$ are trained in advance.

Note that the SVs, $s_{n,f,m}$ for all $f$, depend on the Gaussian index $i^m$ according to (6), (9), and (10), and thus $i^m$ is the frequency independent parameter for the SVs.

### 3.2. Locational model

This section defines the locational model $p(A_{n,f}|m, \theta_A)$. We assume that all sources are located at different locations, and thus have different PDs. We write the PD of each source by $a_{n,f,m}$. From the sparseness assumption (3), we also assume that the PD $a_{n,f,m}$ of the dominant source $m$ is observed as $A_{n,f}$, and other PDs $a_{n,f,m' \neq m}$ are not observed as shown in Fig. 1. That is, a vector $\vec{a}_{n,m}$ is the locational SV and

$$p(A_{n,f}|m, \theta_A) = p(a_{n,f,m} = A_{n,f}|\theta_A).$$

We provide a model for $p(a_{n,f,m}|\theta_A)$ by using a frequency independent mean $\mu_m$ and variance $\sigma^2_m$, where $\mu_m$ corresponds to the expectation value of the TDOA. Moreover, to consider the spatial aliasing issue as in [4], we employ a mixture of wrapped Gaussian distributions [6]. Our locational model is

$$\begin{aligned} p(a_{n,f,m}|\theta_A) &= \sum_{k=-K}^{K} p(a_{n,f,m}|k, \theta_A) p(k) \\ p(a_{n,f,m}|k, \theta_A) &= \frac{1}{\sqrt{2\pi\sigma^2_m}} e^{\left( \frac{-(a_{n,f,m} + 2\pi k - 2\pi f \mu_m)^2}{2\sigma^2_m} \right)}, \end{aligned} \quad (11)$$

where $-\pi \leq a_{n,f,m} < \pi$, the integer $k$ is a phase wrapping index which will be handled as an hidden variable, and $K$ is the maximum wrapping index which can be determined from the microphone spacing and the sampling frequency [4]. In this paper, we disregard the prior $p(\theta_A)$ in (4).

### 3.3. EM algorithm for likelihood maximization

This section provides the optimization algorithm. Let $\theta = \{\mu_m, \sigma_m^2, \vec{\rho}_{n,m}, p(m)\}$ be a model parameter set. In the following, the dominant source index $m$, indices of Gaussians for source log spectra $\{i^m\} = \{i^1, \cdots, i^M\}$, and the phase wrapping index $k$ are dealt with as hidden variables.

The cost function of the MAP estimation is defined based on a log of a joint probability density function as:

$$L(\theta) = \log p(\{\vec{X}_n\}, \{\vec{A}_n\}, \theta)$$
$$= \sum_{n,f} \sum_{m} \sum_{k} \sum_{\{i^m\}} \log p(X_{n,f}, A_{n,f}, m, k | \{i^m\}, \theta) p(\{i^m\}) p(\theta).$$

The above cost function can be maximized by using the EM algorithm. Note that the posterior of the dominant source index $m$ calculated in the E-step decides how the observation vectors are factorized into the SVs. The auxiliary function $Q$ is given as [1]

$$Q = \sum_{n,f} \sum_{m} \sum_{k} \sum_{\{i^m\}} q(m, k, \{i^m\}) \cdot \log \left[ p(X_{n,f}, A_{n,f}, m, k | \{i^m\}, \theta) p(\{i^m\}) p(\theta) \right] \quad (12)$$

$$= \sum_{n,f} \sum_{m} \sum_{k} g_{n,f,m,k} \log p(X_{n,f}, A_{n,f} | m, k, \theta) p(m) p(k) + \sum_{n} \sum_{m'} \sum_{i^{m'}} z_{n,i^{m'}} \log p(\vec{\rho}_{n,m'} | i^{m'}) p(i^{m'}), \quad (13)$$

where $q(m, k, \{i^m\}) = p(m, k, \{i^m\} | X_{n,f}, A_{n,f}, \theta)$ is the posterior

$$q(m, k, \{i^m\}) = g_{n,f,m,k} \prod_{i^m} z_{n,i^m}$$
$$g_{n,f,m,k} = p(m, k | X_{n,f}, A_{n,f}, \theta^t)$$
$$z_{n,i^m} = p(i^m | \vec{\rho}_{n,m}),$$

and we define

$$G_{n,f,m} = p(m | X_{n,f}, A_{n,f}, \theta^t) = \sum_{k} g_{n,f,m,k}. \quad (14)$$

In the above equations, $G_{n,f,m}$ corresponds to the posterior of source $m$ being dominant at a t-f slot $(n, f)$, and $z_{n,i^m}$ is the frequency-independent posterior of $p(i^m)$ which indicates that the $i^m$-th Gaussian component is selected for source $m$ at time slot $n$.

The Q function is maximized by iterating the following E- and M- steps.

**E-step:** The posterior values are calculated in the E-step:

$$g_{n,f,m,k} = \frac{p(X_{n,f} | m, \theta) p(A_{n,f} | m, k, \theta) p(m) p(k)}{\sum_m \sum_k p(X_{n,f} | m, \theta) p(A_{n,f} | m, k, \theta) p(m) p(k)} \quad (15)$$

$$z_{n,i^m} = \frac{p(\vec{\rho}_{n,m} | i^m) p(i^m)}{\sum_{i^m} p(\vec{\rho}_{n,m} | i^m) p(i^m)} \quad (16)$$

**M-step:** In the M-step, the parameter $\theta = \{\mu_m, \sigma_m^2, \rho_{n,f,m}, p(m)\}$ are updated so that the Q function is maximized. The update rules for the locational parameter, $\mu_m$ and $\sigma_m^2$ can be derived simply setting the first derivative of Q with respect to the parameters to zero, and they can be found in [4].

[1] The tips for derivation from (12) to (13) can be found in Appendix.

On the other hand, the update rule for the spectral parameter $\rho_{n,f,m}$ is not such simple, because we are considering the prior (10). Therefore, we discuss how we can update the spectral parameter. When we use (9) and the sparse observation model (7), the Q function related to the parameter $\rho_{n,f,m}$ becomes

$$Q' = \sum_{n,f} \sum_{m} \left( -\frac{G_{n,fm} |X_{n,f}|^2}{\exp(\rho_{n,f,m})} - \rho_{n,f,m} \right) + \sum_{n,f} \sum_{m} \sum_{i^m} z_{n,i^m} \left( -\frac{(\rho_{n,f,m} - \nu_{n,f,i^m})^2}{2 v_{n,f,i^m}} \right)$$

By setting $\frac{\partial Q}{\partial \rho_{n,f,m}^2} = 0$, we obtain a equation:

$$\frac{G_{n,f,m} |X_{n,f}|^2}{\exp(\rho_{n,f,m})} - \rho_{n,f,m} \sum_{i^m} \frac{z_{n,i^m}}{v_{n,f,i^m}} + \left( \sum_{i^m} \frac{z_{n,i^m} \nu_{n,f,i^m}}{v_{n,f,i^m}} - 1 \right) = 0 \quad (17)$$

As this function has a shape of a hinge function with respect to $\rho_{n,f,m}$, the spectral parameter $\rho_{n,f,m}$ can be obtained with the Newton-Raphson method. In our implementation, the Newton-Raphson method converged within three iterations in each E-M update.

The mixture weight $p(m)$ is updated as follows:

$$p(m) = \frac{\sum_{n,f} G_{n,f,m}}{N_n N_f}, \quad (18)$$

where $N_n$ and $N_f$ are the numbers of time frames and frequency bins, respectively. 1

### 3.4. Source Separation

The t-f mask $M_{n,f,m}$ in (2) for the $m$-th separated source is obtained by the posterior (14), that is, $M_{n,f,m} = G_{n,f,m}$. The separated signal is obtained by

$$y_{n,f,m} = G_{n,f,m} x_{n,f,1}. \quad (19)$$

## 4. EXPERIMENTS

### 4.1. Experimental setup

We utilized two microphones whose spacing was 20 cm, that is, the spatial aliasing problem does occur over 850 Hz. The source signals were 5-second Japanese speech signals of two males and two females, sampled at 16 kHz. The number of sources $N_s = 2$ or 3 was always given in the experiments, and the source positions were $45^\circ$ and $150^\circ$ for $N_s = 2$ and $25^\circ$, $90^\circ$ and $150^\circ$ for $N_s = 3$. In this paper, we only considered an anechoic case, i.e., we gave the corresponding delays to the sources to obtain the mixtures. The frame size for STFT was 512 (32 ms), and the frame shift was 128 (8 ms).

The separation performance was evaluated with the signal-to-interference ratio (SIR) as a measure of separation performance, and the signal-to-distortion ratio (SDR) as a measure of sound quality. Their definitions can be found in [9]. We performed six speaker combinations and then averaged the results. The number of the iteration for the EM algorithm was 15.

For the training of the log spectral prior, we utilized 143 sentences for each speaker, where each sentence lasted from 4 to 8 secs. . The number of the Gaussian component $I = 32$. The model was the speaker dependent in this paper.

### 4.2. Results

Figure 2 shows example spectra of the separated signals (A) without and (B) with the spectral model. In Fig. 2 (A), the separation was

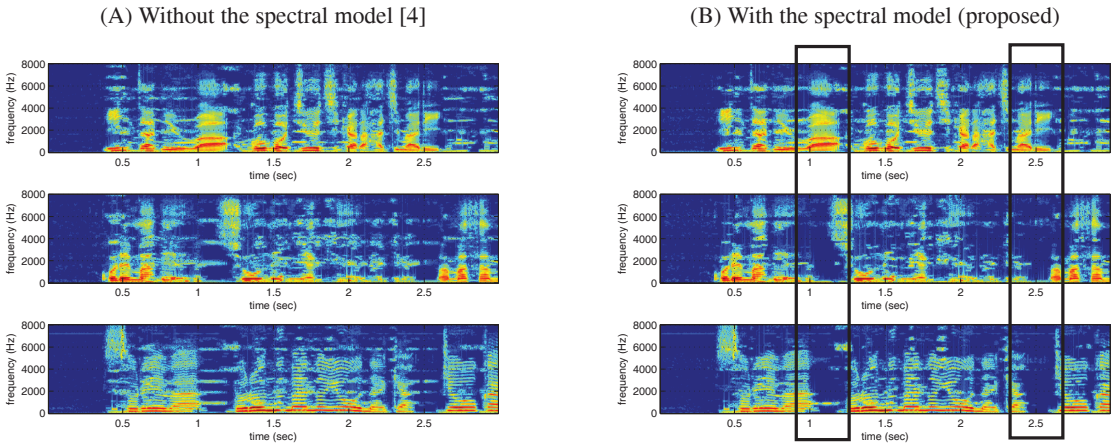


**Fig. 2**. Spectral examples of the separated signals with and without the spectral model ($N_s = 3$).

**Table 1**. Separation results [dB]. The input SIR was 0 [dB] and $-3.0$ [dB] for $N_s = 2$ and $N_s = 3$, respectively.

| | $N_s = 2$ | | $N_s = 3$ | |
|---|---|---|---|---|
| | SIR | SDR | SIR | SDR |
| PD only [4] | 14.89 | 15.81 | 11.80 | 11.87 |
| Naive spectral model [5] | 13.38 | 14.15 | 5.25 | 8.44 |
| Proposed | **15.66** | **15.96** | **12.48** | **12.18** |

achieved with only the PD clustering as in [4]. We can see many horizontal-striped components at the frequencies where the PD of sources lapped over each other due to the spatial aliasing. On the other hand, thanks to the spectral model with the spectral prior, such a horizontal-striped parts are decreased in Fig. 2 (B), especially in the regions indicated by rectangles.

Table 1 summarizes the separation performance with the PD only [4], the naive spectral model (8) [5], and the proposed SVF. With the naive spectral model, performance is insufficient especially for the underdetermined case $N_s = 3$. We observed that separation performance of one of three sources is very poor due to the insufficient solution of the permutation, or sometimes the power of one of three separated signals became very small. The latter comes from the sparse observation model (7), which has high likelihood when only one source has a large variance and the others are zero. On the other hand, by using the proposed method, we can obtain good separation results even for the underdetermined case $N_s = 3$. With the proposed method, we did not observe the phenomenon that one of three separated signals became very small. Moreover, the spectrum at the lapped PDs was refined to the degree of shown in Fig. 2. We can conclude that, the proposed method can achieve the better performance than the previous methods.

## 5. CONCLUSION

We proposed a new approach, the sparse vector factorization (SVF), that inherently prevents the permutation problem. The method utilized the wrapped-phase GMM to model the locational SV, and the source log-spectral model for the spectral SV. In the model, the prior of the log spectral model was also employed. From the results obtained with the simulated experiments, we confirmed that the proposed SVF can separate signals successfully even for an underdetermined case. We also showed that the proposed method improved the separation performance at the frequencies where the PDs of sources overlap due to the spatial aliasing.

Our future study include an evaluation with a speaker independent spectral prior, and the investigation for echoic scenarios, where the observed signals should contain some distortion that is not exist in the spectral prior.

## 6. APPENDIX

The logarithm term in (12) can be written from (4) and (10) as:

$$p(X_{n,f}, A_{n,f}|m,k,\theta)p(m)p(k) \prod_{\{i^m\}} p(\rho_{n,f,m}|i^m)p(\{i^m\}),$$

where $p(X_{n,f}, A_{n,f}|m,\theta)$ is given by (5), (7) and (11), and

$$p(\rho_{n,f,m}|i^m) = N(\rho_{n,f,m}; \nu_{n,f,i^m}, v_{n,f,i^m})$$

(see (10)). By using the posterior in Section 3.3 and the above equations, we can derive (13).

## 7. REFERENCES


[1] O. Yilmaz 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] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. WASPAA2007*, 2007, pp. 147–150.

[3] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. WASPAA2007*, 2007, pp. 139–142.

[4] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in *Proc. ICA'09*, 2009, vol. 5441/2009, pp. 742–750.

[5] S. Araki, T. Nakatani, and H. Sawada, "Sparse source separation based on simultaneous clustering of source locational and spectral features," *Acoustical Science and Technology*, vol. 32, no. 4, pp. 161–164, 2011.

[6] P. Smaragdis and P. Boufounos, "Learning source trajectories using wrapped-phase hidden Markov models," in *Proc. of WASPAA'05*, oct 2005, pp. 114–117.

[7] T. Kristjansson and J. Hershey, "High resolution signal reconstruction," in *Proc. ASRU*, 2003, pp. 291–296.

[8] K. Wilson, "Speech source separation by combining localization cues with mixture models of speech spectra," in *Proc. of ICASSP 2007*, 2007, vol. I, pp. 33–36.

[9] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 77, no. 8, pp. 1833–1847, Aug 2007.