PERMUTATION-FREE CONVOLUTIVE BLIND SOURCE SEPARATION VIA FULL-BAND CLUSTERING BASED ON FREQUENCY-INDEPENDENT SOURCE PRESENCE PRIORS

Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani

NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
2-4, Hikaridai, Seika-cho, "Keihanna Science City" Kyoto 619-0237 Japan

ABSTRACT

We propose permutation-free frequency-domain blind source separation (BSS) via full-band clustering of the time-frequency (T-F) components based on time-varying signal presence priors. Frequency-domain methods of BSS usually process each frequency bin separately, and therefore necessitate the subsequent alignment of the permutation ambiguity that arises between frequency bins. In contrast, the proposed method simultaneously processes all frequency bins by using a mixture model with *time-varying, frequency-independent mixture weights*. We propose to assume non-sparse priors on the mixture weights to prevent the degradation of source separation performance by the time-varying mixture weights. We propose a customized expectation-maximization (EM) algorithm for the maximum *a posteriori* (MAP) estimation of the model parameters, to which we introduce a novel technique to avoid convergence to local maxima. For audio source separation, we use the normalized observation vector as the feature vector, and the Watson mixture model (WMM) as the mixture model. Evaluations confirm that the proposed permutation-free BSS results in source separation performance comparable to the state-of-the-art clustering-based BSS composed of bin-wise clustering and permutation alignment.

***Index Terms*—** Blind source separation, clustering, EM algorithm, mixture model, permutation problem

1. INTRODUCTION

BSS is an important technique with applications such as a front-end for robust automatic speech recognition (ASR). The most standard techniques to BSS include independent component analysis (ICA) [1, 2] and clustering-based approaches [3, 4, 5].

BSS is usually performed in two stages: frequency bin-wise source separation followed by permutation alignment [2, 4, 5]. Among the most standard techniques for permutation alignment is the approach based on direction-of-arrival (DOA) estimates of the bin-wise separated sources [6, 7]. However, DOA estimation is unreliable in low- and high-frequency regions due to the low spatial resolution and spatial aliasing, respectively. Therefore, this approach for permutation alignment may not work well for particular frequencies. Another well-known approach is based on the temporal envelopes of the bin-wise separated sources [7, 8].

In this paper, as an alternative to such two-stage approaches, we propose a one-stage permutation-free BSS method. In the proposed method, the T-F slots are clustered into sources by fitting the normalized observation vector with the WMM in the MAP sense. The EM is employed for deriving efficient update rules. The following distinguishing features enable the avoidance of permutation ambiguity: first, the mixture weights of the WMM are assumed to be dependent on time, and not on the frequency. Second, the mean orientations and the concentration parameters at each frequency are permuted between sources so as to maximize the *a posteriori* probability after each EM iteration. Unlike permutation alignment approaches using the temporal envelope [7, 8], which naturally require certain length of data, the proposed method has the potential of being extended to the online scenario, which is important for adapting to the real-world time-varying environments. Note that the proposed full-band clustering is a general approach in the sense that it may be applied to BSS for any time-varying signals with common temporal modulation.

Note that, in the context of ICA, independent vector analysis (IVA) [9, 10] has been proposed, which can be viewed as the permutation-free extension of ICA. There also exists a full-band clustering-based BSS method [3] based on time-delay of arrival (TDOA). However, TDOA is sensitive to spatial aliasing, so that the clustering by this method requires the array size to be sufficiently small compared to the wavelength. In this paper, we propose full-band permutation-free clustering-based BSS, which can be applied to an arbitrary array.

2. CLUSTERING-BASED BSS USING THE SPATIAL FEATURE

Throughout, we employ the T-F representation and denote by τ and ω the frame index and the angular frequency, respectively. The BSS problem considered in this paper is the estimation of a known number of sources $s_{k\tau\omega} \in \mathbb{C}\,(k = 1, \ldots, K)$ from given M-channel observation $\boldsymbol{y}_{\tau\omega} \in \mathbb{C}^M$. More precisely, $s_{k\tau\omega}$ denotes the kth source image observed at the rth microphone with r denoting the index of the reference microphone, for we are not aiming at dereveberation.

Specifically, we focus on the clustering-based approach. The sparseness property of speech signals is assumed as in [11] such that at any T-F slot, at most one source is active. This results in the following observation model:

$$\boldsymbol{y}_{\tau\omega} = s_{d(\tau,\omega)\tau\omega}\boldsymbol{h}_{d(\tau,\omega)\omega}, \tag{1}$$

where $\boldsymbol{h}_{k\omega}$ denotes the transfer function from the kth source to the microphones, and $d(\tau, \omega)$ the *dominant source index*, that is the index of the source contributing to the T-F slot (τ, ω). Note that the rth entry of $\boldsymbol{h}_{k\omega}$ equals 1 by definition. Under the model as in (1), $s_{k\tau\omega}$ is retrieved through the application of the T-F masks, defined by

$$m_{k\tau\omega} = \begin{cases} 1, & d(\tau,\omega) = k \\ 0, & d(\tau,\omega) \neq k \end{cases}. \tag{2}$$

Indeed,

$$s_{k\tau\omega} = m_{k\tau\omega}[\boldsymbol{y}_{\tau\omega}]_r, \tag{3}$$

where $[\boldsymbol{y}_{\tau\omega}]_r$ denotes the rth entry of $\boldsymbol{y}_{\tau\omega}$. In such clustering-based approaches as in this study, $m_{k\tau\omega}$ is estimated via the clustering of the T-F slots into each source.

We use the *spatial feature* as in [4, 5], which is defined by

$$\boldsymbol{x}_{\tau\omega} \triangleq \frac{\boldsymbol{y}_{\tau\omega}}{\|\boldsymbol{y}_{\tau\omega}\|}, \tag{4}$$

where $\|\cdot\|$ denotes the Euclidean norm. In [4, 5], the spatial features are clustered into sources using a mixture model in a soft manner, by estimating the posterior probability of source presence, $\gamma_{k\tau\omega} \triangleq P(k|\boldsymbol{x}_{\tau\omega}, \Theta)$, where Θ denotes the estimated parameters of the mixture model. The mask (2) can then be designed based on

$$k_{\tau\omega} = \arg\max_k \gamma_{k\tau\omega}. \tag{5}$$

Vu *et al*. [4] proposed the use of the WMM, a mixture model defined on the unit hypersphere, since the features as in (4) are of unit-norm. Sawada *et al*. [5] used a variant of the Gaussian mixture model (GMM) instead, which can be viewed as an approximation of the WMM. However, these methods are performed in each frequency bin separately, requiring post-processing of permutation alignment.

3. PROPOSED PERMUTATION-FREE BSS

3.1. Full-band WMM with frequency-independent time-varying mixture weights

Unlike the frequency-dependent, time-invariant mixture weights in the previous techniques [4, 5], we propose a mixture model with *frequency-independent, time-varying mixture weights* for the clustering of the spatial feature (4). This results in a unified clustering algorithm, which deals with all frequency bins at once. Furthermore, with such mixture weights, a better model fit (*i.e.* higher *a posteriori* probability) is expected, when the permutation is more consistent in all frequency bins.

The proposed WMM is given by

$$p(\boldsymbol{x}_{\tau\omega}|\Theta) = \sum_{k=1}^{K} \alpha_{k\tau} p(\boldsymbol{x}_{\tau\omega}|k, \boldsymbol{a}_{k\omega}, \kappa_{k\omega}), \tag{6}$$

where $\alpha_{k\tau} \triangleq P(k)$ denotes the time-variant frequency-independent mixture weight. The Watson distribution is given by [12]

$$p(\boldsymbol{x}_{\tau\omega}|k, \boldsymbol{a}_{k\omega}, \kappa_{k\omega}) = \frac{(M-1)!}{2\pi^M M(1, M, \kappa_{k\omega})} \exp(\kappa_{k\omega} |\boldsymbol{a}_{k\omega}^{\mathsf{H}} \boldsymbol{x}_{\tau\omega}|^2), \tag{7}$$

where $\boldsymbol{a}_{k\omega}$ is called the mean orientation, $\kappa_{k\omega}$ the concentration parameter, $M(a, b, x)$ is the Kummer function [13]. The parameter set is given as:

$$\Theta \triangleq \{\{\alpha_{k\tau}\}_{k\tau}, \{\boldsymbol{a}_{k\omega}\}_{k\omega}, \{\kappa_{k\omega}\}_{k\omega}\}. \tag{8}$$

In order to control the degree to which $\alpha_{k\tau}$ affects the source separation result, we employ a Dirichlet prior for $\alpha_{k\tau}$ as follows:

$$p(\{\alpha_{k\tau}\}_k) = \frac{\Gamma(K\phi)}{\Gamma(\phi)^K} \prod_{k=1}^{K} \alpha_{k\tau}^{\phi-1}, \tag{9}$$

where Γ is the gamma function. The larger the value of ϕ, the less the effect of $\alpha_{k\tau}$ becomes. We investigate the effect of ϕ on source separation performance in Section 4. We assume uniform priors for the parameters other than $\{\alpha_{k\tau}\}_{k\tau}$. Therefore, $p(\Theta) = \prod_\tau p(\{\alpha_{k\tau}\}_k)$.

3.2. MAP estimation of the parameters via EM

We employ the *a posteriori* probability as the objective function. Under the assumption that $\{\boldsymbol{x}_{\tau\omega}\}_{\tau\omega}$ are independent from one another, this probability is given by the following equation:

$$\begin{aligned}\log p(\Theta|\{\boldsymbol{x}_{\tau\omega}\}_{\tau\omega}) &= \sum_{\tau\omega} \log p(\boldsymbol{x}_{\tau\omega}|\Theta) + \log p(\Theta) \\ &= \sum_{\tau\omega} \log \sum_{k=1}^{K} \alpha_{k\tau} p(\boldsymbol{x}_{\tau\omega}|k, \boldsymbol{a}_{k\omega}, \kappa_{k\omega}) \\ &\quad + (\phi - 1) \sum_{k\tau} \log \alpha_{k\tau},\end{aligned} \tag{10}$$

where the equations hold up to a constant independent of Θ. This should be maximized subject to the following constraints:

$$\sum_{k=1}^{K} \alpha_{k\tau} = 1, \|\boldsymbol{a}_{k\omega}\| = 1. \tag{11}$$

Although (10) is a nonlinear function with a summation in logarithm, we can derive an efficient iterative algorithm based on the EM, by viewing k as a hidden variable. The E-step amounts to the estimation of the posterior probability $\gamma_{k\tau\omega}$ using the current parameter estimate Θ'. Considering the Bayes rule, this is performed as follows:

$$\gamma_{k\tau\omega} = \frac{\alpha'_{k\tau} p(\boldsymbol{x}_{\tau\omega}|k, \boldsymbol{a}'_{k\omega}, \kappa'_{k\omega})}{\sum_{l=1}^{K} \alpha'_{l\tau} p(\boldsymbol{x}_{\tau\omega}|l, \boldsymbol{a}'_{l\omega}, \kappa'_{l\omega})}. \tag{12}$$

In the M-step, the following Q function, defined using the updated posteriors $\gamma_{k\tau\omega}$, is maximized with respect to each parameter:

$$\begin{aligned}Q(\Theta, \Theta') &= \sum_{k\tau\omega} \gamma_{k\tau\omega} \log[\alpha_{k\tau} p(\boldsymbol{x}_{\tau\omega}|k, \boldsymbol{a}_{k\omega}, \kappa_{k\omega})] \qquad (13) \\ &\quad + (\phi - 1) \sum_{k\tau} \log \alpha_{k\tau} \\ &= \sum_{k\tau} \left[\sum_{\omega} \gamma_{k\tau\omega} + (\phi - 1)\right] \log \alpha_{k\tau} \qquad (14) \\ &\quad - \sum_{k\omega} \left[\sum_{\tau} \gamma_{k\tau\omega}\right] \log M(1, M, \kappa_{k\omega}) \\ &\quad + \sum_{k\omega} \left[\sum_{\tau} \gamma_{k\tau\omega}\right] \kappa_{k\omega} \boldsymbol{a}_{k\omega}^{\mathsf{H}} \boldsymbol{R}_{k\omega} \boldsymbol{a}_{k\omega},\end{aligned}$$

where (14) holds up to a constant independent of Θ, and

$$\boldsymbol{R}_{k\omega} \triangleq \frac{\sum_\tau \gamma_{k\tau\omega} \boldsymbol{x}_{\tau\omega} \boldsymbol{x}_{\tau\omega}^{\mathsf{H}}}{\sum_\tau \gamma_{k\tau\omega}}. \tag{15}$$

The update rules are derived through the differentiation of (14) w.r.t. each parameter, with the constraints in (11) introduced using Lagrangian multipliers. Due to space limitation, here we present the results only. $\boldsymbol{a}_{k\omega}$ is updated as a normalized principal eigenvector of $\boldsymbol{R}_{k\omega}$. $\alpha_{k\tau}$ is updated by

$$\alpha_{k\tau} = \frac{\sum_\omega \gamma_{k\tau\omega} + (\phi - 1)}{F + (\phi - 1)K}, \tag{16}$$

where F denotes the number of frequency bins. We see that, when $\phi = 1$, $\alpha_{k\tau}$ is updated as the average of $\gamma_{k\tau\omega}$ over frequencies.

With an increasing value of ϕ, $\alpha_{k\tau}$ approaches to the constant $1/K$. Regarding $\kappa_{k\omega}$, we have the following equation:

$$\frac{M'(1, M, \kappa_{k\omega})}{M(1, M, \kappa_{k\omega})} = \boldsymbol{a}_{k\omega}^{\mathsf{H}} \boldsymbol{R}_{k\omega} \boldsymbol{a}_{k\omega} = \lambda_{k\omega}, \tag{17}$$

where $M'(a, b, x) \triangleq \frac{\partial}{\partial x} M(a, b, x)$, and $\lambda_{k\omega}$ is the principal eigenvalue of $\boldsymbol{R}_{k\omega}$. (17) can be approximately solved as follows [13]:

$$\kappa_{k\omega} \sim \frac{M\lambda_{k\omega} - 1}{2\lambda_{k\omega}(1 - \lambda_{k\omega})} \left[1 + \sqrt{1 + \frac{4(M+1)\lambda_{k\omega}(1 - \lambda_{k\omega})}{M - 1}} \right]. \tag{18}$$

Note that the spatial feature is prewhitened as in [5]. The normalization is performed after the prewhitening again.

3.3. Parameter permutation procedure

Unfortunately, the EM algorithm presented in Section 3.2 as it is tends to converge to a nonoptimal local maximum of the objective function (10), which is also a permutated solution.

The result of a preliminary experiment is presented here to illustrate this issue. We compared the parameters estimated by the EM algorithm for $K = 2$ sources for the two different ways of initialization of the mean orientations, namely the oracle initialization calculated from the known source images, and the permuted initialization obtained by permuting the mean orientations for the oracle initialization between 1 and 2 kHz. The other parameters were initialized in a manner described in Section 4 for both cases. Fig. 3.2 plots the phase difference between microphones of the estimated mean orientations as a function of the frequency for (a) the oracle and (b) the permuted initializations and the EM algorithm with (c) the oracle or (d) the permuted initializations. Note that we plotted only the phase difference, not the amplitude ratio, for the sake of easy interpretation. For both ways of initialization, the result of the EM algorithm looks similar to the initialization. The objective *a posteriori* probability was 70714 and 56909 for (c) and (d), respectively. These facts, combined together, implies that both (c) and (d) are local maxima of the objective function, but the (c) corresponds to a higher local maximum. This motivates us to think of a way of avoiding local maxima.

In order to avoid convergence to permuted solutions, we permute the mean orientation and the concentration parameters at each frequency after each EM iteration, so that the likelihood is maximized as:

$$\Pi \leftarrow \arg\max_{\Pi} \sum_{\tau} \log \sum_{k=1}^{K} \alpha'_{k\tau} p(\boldsymbol{x}_{\tau\omega} | \boldsymbol{a}'_{\Pi(k)\omega}, \kappa'_{\Pi(k)\omega}), \tag{19}$$

$$\boldsymbol{a}'_{k\omega} \leftarrow \boldsymbol{a}'_{\Pi(k)\omega}, \tag{20}$$

$$\kappa'_{k\omega} \leftarrow \kappa'_{\Pi(k)\omega}, \tag{21}$$

where $\Pi : \{1, 2, \ldots, K\} \rightarrow \{1, 2, \ldots, K\}$. The nondecrease of the likelihood function through this modified EM iteration is guaranteed. Figs. 3.2(e) and (f) show the result of this modified EM algorithm with the oracle and the permuted initializations. We see that the results look very similar to the ideal solution (a) for both initializations. Actually, the objective *a posteriori* probability was 81874 for both cases.

4. EXPERIMENTS

To demonstrate that the proposed full-band clustering works effectively without any additional permutation alignment, we compared the performance of the following BSS methods:

Table 1. Experimental conditions

Number of microphones	$M = 2$
Number of sources	$K = 2$
Source signals	speeches of 8 s
Reverberation time	$\mathrm{RT}_{60} = 130$-440 ms
Sampling rate	8 kHz
STFT frame size	1024 (128 ms)
STFT frame shift	256 (32 ms)

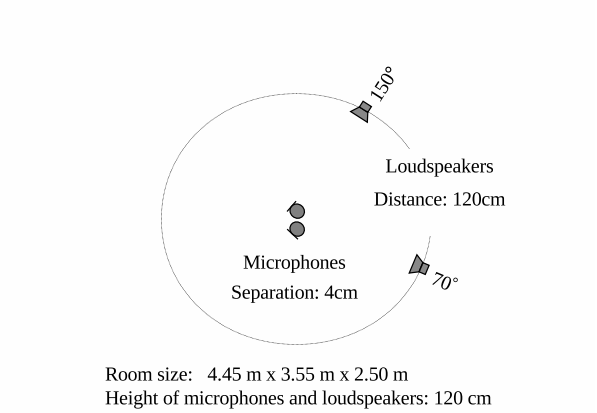

Fig. 2. Configuration of the microphones and the sources in the experiments.

- Bin-wise clustering followed by an established permutation alignment method [5]. The bin-wise clustering is realized by choosing $\phi = \infty$ in the proposed algorithm, which makes the mixture weight constant: $\alpha_{k\tau} = 1/3$.
- Proposed full-band clustering for $\phi = 1, 10, 10^2, 10^3$. Here $\phi = 1$ corresponds to the maximum-likelihood (ML) solution without the Dirichlet prior.

The effectiveness of the proposed method can be confirmed if the achieved separation performance is comparable to the bin-wise clustering followed by the permutation alignment for some ϕ.

The experimental conditions are summarized in Table 1. We generated the mixture by convolving the clean sources with measured impulse responses, and by summing up thus generated source images. The geometrical configuration used for the recording of the impulse responses is described in Fig. 2. We evaluated the separation performance by signal-to-distortion ratio (SDR) as defined in [14]. To alleviate the variation of performance depending on speech samples, we averaged SDR values for 8 speech combinations. The EM algorithm is initialized as follows as in [4, 5]: $\alpha_{k\tau} = \frac{1}{K}$, $\kappa_{k\omega} = 20$, and $\boldsymbol{a}_{k\omega}$ by choosing randomly from $\{\boldsymbol{x}_{\tau\omega}\}_{\tau\omega}$. The number of EM iterations was fixed to 100.

Table 2 plots the SDR results for the compared methods as a function of the reverberation time. We see that the proposed method for $\phi = 100$ and $\phi = 1000$ gave SDR comparable to the bin-wise clustering followed by the permutation alignment.

5. CONCLUSION

This paper proposed a permutation-free BSS method based on the clustering of the normalized observation vector using the WMM

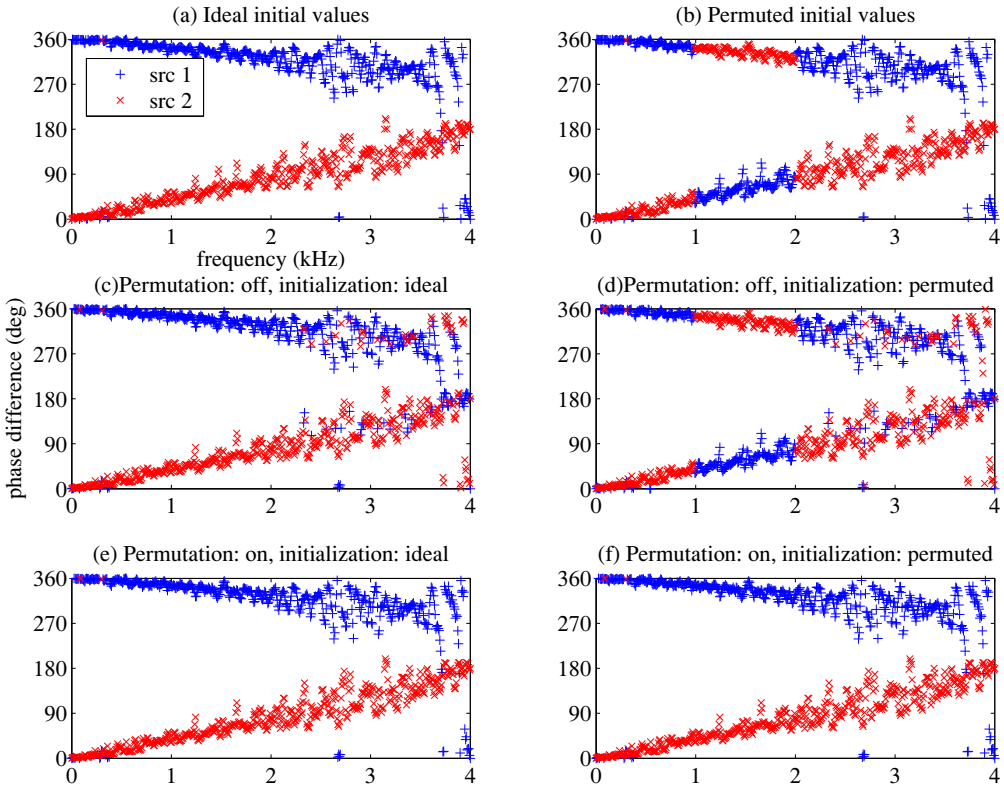

Fig. 1. The phase difference between the microphones of the estimated mean orientation of each cluster as a function of the frequency.

Table 2. SDR for the proposed full-band approach and the bin-wise approach followed by permutation alignment for varying reverberation times RT_{60}. The figures in the parenthesis represent the value of the hyperparameter ϕ.

RT_{60} (ms)	130	200	250	300	370	440
Bin-wise	**16.1**	13.6	11.3	10.5	10.0	**8.8**
Proposed (1)	15.3	12.9	10.4	9.6	9.2	8.0
Proposed (10)	15.5	13.1	10.6	9.8	9.4	8.2
Proposed (10^2)	15.8	13.5	11.4	10.7	9.9	8.6
Proposed (10^3)	**16.1**	**13.7**	**11.7**	**10.6**	**10.1**	8.7

with time-varying frequency-independent mixture weights. The clustering is performed via MAP estimation of the parameters of the WMM. A procedure of permuting parameters between sources at each EM iteration is employed in order to avoid the nonoptimal convergence of the parameters to permutated solutions. The experiments showed that, by properly choosing the predetermined hyperparameter, the proposed full-band clustering method yields source separation performance comparable to that of the bin-wise clustering followed by subsequent permutation alignment.

The future work includes the extension of the approach to the case of unknown (and possibly time-varying) number of sources, and to the online BSS.

6. REFERENCES

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

[2] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds., pp. 299–327. Springer, Berlin, 2005.

[3] 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.

[4] D. H. T. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. ICASSP*, Mar. 2010, pp. 241–244.

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

[6] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain

blind speech separation," in *Proc. ICASSP*, May 2002, pp. 881–884.

[7] 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. ASLP*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

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

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

[10] T. Kim, T. Eltoft, and T. W. Lee, "Independent vector analysis: an extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.

[11] A. Jourjine, S. Rickard, and Ö. Yılmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," in *Proc. ICASSP*, Jun. 2000, pp. 2985–2988.

[12] K. V. Mardia and I. L. Dryden, "The complex Watson distribution and shape analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 913–926, 1999.

[13] S. Sra and D. Karp, "The multivariate Watson distribution: maximum-likelihood estimation and other aspects," 2012, arXiv: 1104.4422v2.

[14] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Proc. ICA*, 2007, pp. 552–559.