Comparison of superimposition and sparse models in blind source separation by multichannel Wiener filter

Ryutaro Sakanashi*, Shigeki Miyabe†, Takeshi Yamada‡ and Shoji Makino§
*Graduate School of Systems and Information Engineering, University of Tsukuba, Japan
†§Life Science Center of Tsukuba Advanced Research Alliance, University of Tsukuba, Japan
†‡§Faculty of Engineering, Information and Systems, University of Tsukuba, Japan
E-mail: *sakanashi@mmlab.cs.tsukuba.ac.jp, {†miyabe,§maki}@tara.tsukuba.ac.jp, ‡takeshi@cs.tsukuba.ac.jp

***Abstract*—Multichannel Wiener filter proposed by Duong *et al.* can conduct underdetermined blind source separation (BSS) with low distortion. This method assumes that the observed signal is the superimposition of the multichannel source images generated from multivariate normal distributions. The covariance matrix in each time-frequency slot is estimated by an EM algorithm which treats the source images as the hidden variables. Using the estimated parameters, the source images are separated as the maximum a posteriori estimate. It is worth nothing that this method does not assume the sparseness of sources, which is usually assumed in underdetermined BSS. In this paper we investigate the effectiveness of the three attributes of Duong's method, i.e., the source image model with multivariate normal distribution, the observation model without sparseness assumption, and the source separation by multichannel Wiener filter. We newly formulate three BSS methods with the similar source image model and the different observation model assuming sparseness, and we compare them with Duong's method and the conventional binary masking. Experimental results confirmed the effectiveness of all the three attributes of Duong's method.**

I. Introduction

Source separation is an essential technology for hands-free speech recognition and understanding of the sound environment by computer in real environments. In past decades there has been a rapid progress in the research field of blind source separation (BSS), which does not require prior information, such as source position and voice activity detection [1]. Independent component analysis [2] is representative of BSS, but this method cannot be applied to underdetermined conditions where the number of sources is larger than that of microphones. Since stereo recording devices are used widely, underdetermined BSS technique is demanded so that we can apply BSS to two-channel recordings of arbitrary number of sources.

A typical approach to underdetermined BSS is to assume *sparseness* among sources [3]. Sparseness is the nature of signals whose energy is concentrated in some frequencies and almost zero in the other frequencies at each short time period. Thus the BSS methods of this type assume that each time-frequency slot of speech mixture is dominated by one of the sources. The most representative of those methods is time-frequency masking [3], which identifies the dominant source in each time-frequency slot, and separate each source by masking the time-frequency slots where the other sources are dominant. Such identification is done by clustering the multichannel features such as inter-channel level or phase differences with k-means clustering [4], EM algorithm [5], etc.

There have been various works related with the time-frequency masking. Araki *et al.* proposed combination with minimum variance distortionless response beamformer [6] to reduce musical noise caused by masking [7]. Also, Iso *et al.* [8] used the time-frequency binary masking [9] for initialization of an initialization-sensitive BSS method using multichannel Wiener filter proposed by Duong *et al.* [10], whose details are discussed in this paper.

The problem of the time-frequency binary masking is degradation of speech quality caused by modeling mismatch of the sparseness assumption, which is satisfied approximately but not perfectly. While most of the underdetermined BSS techniques assume sparseness of speech, Duong *et al.* proposed low-distortion underdetermined BSS using the multichannel Wiener filter without the sparseness assumption [10]. This method models source image, which is the multiplication of each source and its related transfer functions. The source image is modeled by multivariate normal distribution whose parameterization assumes stationarity of the transfer system and quasi-stationarity of the sources. The multichannel Wiener filter is designed using the parameters estimated by an EM algorithm. The details of the underlying probabilistic model is not described in the original paper [10], but as pointed in [11], the source images are dealt as hidden variables, and the sum of the hidden source images are constrained to equal to the observed signal. With this constraint, superimposition of the multiple sources in the mixed signals is modeled effectively. The obtained multichannel Wiener filtering corresponds to maximum a priori (MAP) estimate of the source images. The detailed probabilistic modeling is described in this paper.

As described above, Duong's method has various interesting attributes. However, it has not been well-investigated which factor is effective for the high-quality source separation. In this paper, we analyze in detail some of these attributes which are considered to be the reason why Duong's method can achieve

high quality source separation. In particular, we focus attention on the following three, i.e., 1) the source image model by multivariate normal distribution, 2) source separation by the multichannel Wiener filter, and 3) use of EM algorithm developed without assumption of sparseness. In order to confirm the effectiveness of these factors, we formulate alternative underdetermined BSS methods with the same source image model as Duong's method by multivariate normal distribution and the different observation model of sparse source image occurrences but superimposition of multiple source images. We design the following three BSS schemes using the estimated parameters: A) separation using the posterior probability that source image is active, B) separation by maximum likelihood binary mask which corresponds to maximum a posteriori estimation of sparse source images, and C) separation by an alternative multichannel Wiener filter designed using the expectation of the sources in the sparse model. We compared these three methods with Duong's method and MENUET which is a general binary masking method, and discuss the performances from the following three viewpoints; i) validity of the source image model with multivariate normal distribution by comparing the typical binary mask MENUET and sparse binary mask formulated, ii) effectiveness of separation by the multichannel Wiener filter by comparing the above three methods A), B), and C), and iii) effectiveness of source image superposition model not assuming the sparseness by comparing the newly formulated sparse Wiener filter and Duong's method. As a result, effectiveness of all the above three attributes of Duong's method has been confirmed.

In Section 2, we describe the observation model dealt in this paper. We review Duong's method in Section 3. The sparse model which is a comparative approach is formulated in Section 4. In Section 5, we perform a comparison experiment of source separation to confirm the validity of Duong's method. Finally, Section 6 concludes this paper.

II. Observation Model

Here we describe the observation model of Duong's method and sparse models we newly formulate in this paper. First, the observed signal is expressed as

$$\begin{aligned}\mathbf{x}(n,f) &= [x_1(n,f),...,x_I(n,f)]^T \\ &\approx \sum_{j=1}^{J} \mathbf{h}_j(f) s_j(n,f), \end{aligned} \tag{1}$$

$$\mathbf{h}_j(f) = [h_{1j}(f),...,h_{Ij}(f)]^T. \tag{2}$$

where $s_j(n,f)$ is the source signal, $\mathbf{h}_j(f)$ is the transfer function vector of the j-th source whose entry h_{ij} represents the transfer function from the j-th source to the i-th microphone, J is the number of the sources, I is the number of the microphones, and $(\cdot)^T$ is the transpose. Also, we define the component of the j-th source reaching at the microphones as the source image $\mathbf{c}_j(n,f)$, which contains spatial information such as reverberation, expressed as

$$\mathbf{c}_j(n,f) = \mathbf{h}_j(f) s_j(n,f). \tag{3}$$

In this paper, we discuss the problem to estimate the source image $\mathbf{c}_j(n,f)$ of each source.

III. Multichannel Wiener Filter By Duong *et al*.

Here we explain about the multichannel Wiener filter by Duong *et al*., from the viewpoint of maximum likelihood estimation problem, which was not explicitly formulated in [10].

A. Probabilistic Modeling of Source Image

Originally, the source image can be expressed as (3), but assuming the time variation of the transfer function effected by the longer reverberation than the length of the analysis frames, Eq. (3) is rewritten as

$$\mathbf{c}_j(n,f) = \mathbf{h}_j(n,f) s_j(n,f), \tag{4}$$

where $\mathbf{h}_j(n,f)$ is defined as the time-varying transfer function. For the probabilistic modeling of the source image $\mathbf{c}_j(n,f)$, two assumptions are introduced. First, under the assumption that the sources does not change the position, we regard the spatial correlation to be stationary and each of the j-th source is given the time-invariant spatial correlation matrix $\mathbf{R}_j(f)$. Second, assuming speech is non-stationary but quasi-stationary, we denote the time-varying variance of the j-th source with $\nu_j(n,f)$. Under these assumptions, the covariance matrix $\mathbf{R}_{\mathbf{c}_j}(n,f)$ of $\mathbf{c}_j(n,f)$ is regarded as the product of $\mathbf{R}_j(f)$ and $\nu_j(n,f)$ as

$$\mathbf{R}_{\mathbf{c}_j}(n,f) = \nu_j(n,f)\mathbf{R}_j(f), \tag{5}$$

and the source image $\mathbf{c}_j(n,f)$ is assumed to be generated from the zero-mean multivariate normal distribution $\mathcal{N}_{\mathrm{c}}(\mathbf{c}_j(n,f); \mathbf{0}, \mathbf{R}_{\mathbf{c}_j}(n,f))$:

$$p(\mathbf{c}_j(n,f)|\theta(f)) = \mathcal{N}_{\mathrm{c}}(\mathbf{c}_j(n,f); \mathbf{0}, \mathbf{R}_{\mathbf{c}_j}(n,f)). \tag{6}$$

Here the probability density function $\mathcal{N}_{\mathrm{c}}(\mathbf{c}; \mu, \mathbf{\Sigma})$ of the I-dimensional random variable $\mathbf{c}$ arose from the multivariate normal distribution with the mean μ and the covariance matrix $\mathbf{\Sigma}$ is given by

$$\begin{aligned} p(\mathbf{c}; \mu, \mathbf{\Sigma}) &= \mathcal{N}_{\mathrm{c}}(\mathbf{c}; \mu, \mathbf{\Sigma}) \\ &\triangleq \frac{1}{\pi^I \det(\mathbf{\Sigma})} \exp\left(-(\mathbf{c}-\mu)^H \mathbf{\Sigma}^{-1}(\mathbf{c}-\mu)\right), \end{aligned} \tag{7}$$

where $\det(\cdot)$ is the determinant of the square matrix, and $(\cdot)^H$ is the complex conjugate transpose. Note that the spatial correlation matrix $\mathbf{R}_j(f)$ should be expressed as $\mathbf{R}_j(f) = \mathbf{h}_j(f)\mathbf{h}_j^H(f)$ with its rank one when time-invariant transfer function $\mathbf{h}_j(n,f)$ is assumed. However, considering the effects of such long reverberation which does not fit in the analysis frames, $\mathbf{R}_j(f)$ is assumed to be full-rank considering the time variation of the transfer function vector $\mathbf{h}_j(n,f)$. Maximum likelihood estimation of the generative model of $\mathbf{c}_j(n,f)$ and $\mathbf{x}(n,f)$ is described in the following section.

B. Derivation of EM Algorithm

Assuming the observed signal $\mathbf{x}(n,f)$ appears as the superimposition of the source images $\mathbf{c}_j(n,f)$, $j=1,...,J$ as

$$\begin{aligned}\mathbf{x}(n,f) &= \sum_{j=1}^{J}\mathbf{c}_j(n,f)\\ \Leftrightarrow \mathbf{c}_j(n,f) &= \mathbf{x}(n,f) - \sum_{j=1}^{J-1}\mathbf{c}_j(n,f),\end{aligned} \tag{8}$$

the joint probability of $\mathbf{x}(n,f)$ and $\mathbf{c}_j(n,f)$ is derived from Eqs. (7) and (8) as

$$\begin{aligned}&p(\mathbf{C}(n,f),\mathbf{x}(n,f)|\theta(f))\\ &= \prod_{j=1}^{J-1}\mathcal{N}_{\mathrm{c}}\left(\mathbf{c}_j(n,f);\mathbf{0},\mathbf{R}_{\mathbf{c}_j}(n,f)\right)\\ &\quad\cdot\mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f)-\sum_{j=1}^{J-1}\mathbf{c}_j(n,f);\mathbf{0},\mathbf{R}_{\mathbf{c}_J}(n,f)\right),\end{aligned} \tag{9}$$

where $\mathbf{C}(n,f)=\{\mathbf{c}_1(n,f),...,\mathbf{c}_j(n,f),...,\mathbf{c}_J(n,f)\}$. Moreover the likelihood of the observation is obtained as

$$\begin{aligned}p(\mathbf{x}(n,f)|\theta(f)) &= \int p(\mathbf{C}(n,f),\mathbf{x}(n,f)|\theta(f))d\mathbf{C}_J(n,f)\\ &= \mathcal{N}_{\mathrm{c}}(\mathbf{x}(n,f);\mathbf{0},\mathbf{R}_{\mathbf{x}}(n,f)),\end{aligned} \tag{10}$$

by marginalization of all the source images, where $\mathbf{C}_J(n,f)=\{\mathbf{c}_1(n,f),...,\mathbf{c}_{J-1}(n,f)\}$, and $\mathbf{R}_{\mathbf{x}}(n,f)$ is covariance matrix of observation matrix $\mathbf{x}(n,f)$, given as

$$\mathbf{R}_{\mathbf{x}}(n,f)=\sum_{j=1}^{J}\nu_j(n,f)\mathbf{R}_j(f). \tag{11}$$

With these probability density functions, the Q-function is formulated as

$$\begin{aligned}&Q(\theta(f),\bar{\theta}(f))\\ &= \sum_{n=1}^{N}\int p\left(\mathbf{C}(n,f)|\mathbf{x}(n,f),\theta(f)\right)\\ &\qquad\cdot\log p\left(\mathbf{C}(n,f),\mathbf{x}(n,f)|\bar{\theta}(f)\right)d\mathbf{C}_J(n,f)\\ &= \sum_{n=1}^{N}\int\frac{p(\mathbf{C}(n,f),\mathbf{x}(n,f)|\theta(f))}{p\left(\mathbf{x}(n,f)|\theta(f)\right)}\\ &\qquad\cdot\log p\left(\mathbf{C}(n,f),\mathbf{x}(n,f)|\bar{\theta}(f)\right)d\mathbf{C}_J(n,f)\\ &= \sum_{n=1}^{N}\Bigg(-IJ\log\pi - I\sum_{j=1}^{J}\log\bar{\nu}_j(n,f)\\ &\qquad-\sum_{j=1}^{J}\log\det(\bar{\mathbf{R}}_j(f))\\ &\qquad-\sum_{j=1}^{J}\frac{1}{\bar{\nu}_j(n,f)}\mathrm{Tr}\left(\mathbf{M}_j(n,f)\bar{\mathbf{R}}_j^{-1}(f)\right)\Bigg),\end{aligned} \tag{12}$$

$$\begin{aligned}&\mathbf{M}_j(n,f)\\ &= \mathbf{R}_{\mathbf{c}_j}(n,f)-\mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{R}_{\mathbf{c}_j}(n,f)\\ &\quad+\mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f)\mathbf{x}^H(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{R}_{\mathbf{c}_j}(n,f).\end{aligned} \tag{13}$$

The EM algorithm to estimate the parameters is obtained by setting the partial differential of the Q-function to be zero.

C. Parameter Estimation and Source Separation

First, set appropriate initial values of $\nu_j(n,f)$ and $\mathbf{R}_j(f)$, and initialize $\mathbf{R}_{\mathbf{c}_j}(n,f)$ and $\mathbf{R}_{\mathbf{x}}(n,f)$ as follows.

$$\mathbf{R}_{\mathbf{c}_j}(n,f)=\nu_j(n,f)\mathbf{R}_j(f), \tag{14}$$

$$\mathbf{R}_{\mathbf{x}}(n,f)=\sum_{j=1}^{J}\mathbf{R}_{\mathbf{c}_j}(n,f). \tag{15}$$

After the above initialization, the E-step estimates the a posteriori covariance matrix $\mathbf{R}_{\mathbf{c}_j}(n,f)$ as

$$\mathbf{W}_j(n,f)=\mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f), \tag{16}$$

$$\hat{\mathbf{c}}_j(n,f)=\mathbf{W}_j(n,f)\mathbf{x}(n,f), \tag{17}$$

$$\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)=\hat{\mathbf{c}}_j(n,f)\hat{\mathbf{c}}_j^H(n,f)+(\mathbf{I}-\mathbf{W}_j(n,f))\mathbf{R}_{\mathbf{c}_j}(n,f). \tag{18}$$

The update of the M-step is given by

$$\nu_j(n,f)=\frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)), \tag{19}$$

$$\mathbf{R}_j(f)=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{\nu_j(n,f)}\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f), \tag{20}$$

$$\mathbf{R}_{\mathbf{c}_j}(n,f)=\nu_j(n,f)\mathbf{R}_j(f), \tag{21}$$

$$\mathbf{R}_{\mathbf{x}}(n,f)=\sum_{j=1}^{J}\nu_j(n,f)\mathbf{R}_j(f), \tag{22}$$

where $\mathbf{I}$ is the $I\times I$ identity matrix. After the convergence of the iteration of the E- step and M-step, the source image is estimated by using the parameters. Marginalization of $p(\mathbf{C}(n,f)|\mathbf{x}(n,f),\theta(f))$ about $\mathbf{c}_1(n,f)$, $...,\mathbf{c}_{j-1}(n,f),\mathbf{c}_{j+1}(n,f),...,\mathbf{c}_J(n,f)$ gives

$$\begin{aligned}&p(\mathbf{c}_j(n,f)|\mathbf{x}(n,f),\theta(f))\\ &= \mathcal{N}_{\mathrm{c}}\Bigg(\mathbf{c}_j(n,f);\mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f),\\ &\qquad\left(\mathbf{R}_{\mathbf{c}_j}^{-1}(n,f)+\left(\mathbf{R}_{\mathbf{x}}(n,f)-\mathbf{R}_{\mathbf{c}_j}(n,f)\right)^{-1}\right)^{-1}\Bigg),\end{aligned} \tag{23}$$

and each source image can be estimated by the multichannel Wiener filter represented as

$$\mathbf{c}_j(n,f)=\mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f), \tag{24}$$

which corresponds both to the expectation and the MAP estimate.

IV. Formulation Of The Separation Method And Observation Model By Assumption Of Sparseness

In this section, for the comparison to verify the effectiveness of the model of Duong's method, we formulate new source separation methods which have the similar source image model to Duong's method but with an alternative assumption that the observation is generated from the sparse distribution of the source images.

A. Problem Establishment

We assume that in each time-frequency slot only one source is active, and we denote the index of the active source in the time-frequency slot (n, f) as $z(n, f)$. We define the prior probability of the j-th source to be active as

$$p(z(n,f)=j)=\mu_j(f),\ \sum_{j=1}^{J}\mu_j(f)=1. \tag{25}$$

Under this assumption, the observed signal $\mathbf{x}(n, f)$ is modeled as

$$\mathbf{x}(n,f)=\mathbf{c}_{z(n,f)}(n,f), \tag{26}$$

say, the sparse observation model. Similarly to Duong's method, the generation of the source image $\mathbf{c}_j(n, f)$ of this model is represented as

$$\begin{aligned}&p\left(\mathbf{c}_j(n,f)|z(n,f)=j,\theta(f)\right)\\&=\mathcal{N}_{\mathrm{c}}\left(\mathbf{c}_j\left(n,f\right);\mathbf{0},\nu_j\left(n,f\right)\mathbf{R}_j(f)\right),\end{aligned} \tag{27}$$

by assuming that the covariance matrix $\mathbf{R}_{\mathbf{c}_j}(n, f)$ is given by the product of the time-varying variances $\nu_j(n, f)$ and the time-invariant spatial correlation matrix $\mathbf{R}_j(f)$, and the source image $\mathbf{c}_j(n, f)$ follows the normal distribution $\mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f);\mathbf{0},\nu_j(n,f)\mathbf{R}_j(f)\right)$. We represent the model parameters by $\theta(f)=\{\mathbf{R}_j(f),\nu_j(n,f),\mu_j(f),\text{for } j=1,...,J\}$.

B. Model Parameter Estimation

Likelihood $\prod_n p(\mathbf{x}(n,f)|\theta(f))$ of the observed signal $\mathbf{x}(n, f)$ to maximize the model parameters $\theta(f)$ are defined as

$$\begin{aligned}&\prod_n p\left(\mathbf{x}(n,f)|\theta(f)\right)\\&=\prod_n\sum_{j=1}^{J}\mu_j(f)\mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f);\mathbf{0},\nu_j(n,f)\mathbf{R}_j(f)\right).\end{aligned} \tag{28}$$

This maximum likelihood estimation can be obtained by EM algorithm to maximize the Q-function with the hidden variable $z(n, f)$ is given below.

$$\begin{aligned}&Q_f(\theta(f),\bar{\theta}(f))\\&=\sum_{n,j}m_j(n,f)\log\mu_j(f)\mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f);\mathbf{0},\nu_j(n,f)\mathbf{R}_j(f)\right)\\&=\sum_{n,j}m_j(n,f)\Bigg(\log\mu_j(f)-I\log\pi-I\log\nu_j(n,f)\\&\quad-\log\det\left(\bar{\mathbf{R}}_j(f)\right)-\frac{\mathbf{x}(n,f)^{\mathrm{H}}\bar{\mathbf{R}}_j^{-1}(f)\mathbf{x}(n,f)}{\nu_j(n,f)}\Bigg),\end{aligned} \tag{29}$$

where $m_j(n, f)$ behaving as the time-frequency soft mask is the posterior probability of $z(n, f)$ given as

$$m_j(n,f)=p\left(z(n,f)=j|\mathbf{x}(n,f),\mathbf{R}_j(f)\nu_j(n,f)\right), \tag{30}$$

and satisfies

$$\sum_{j=1}^{J}m_j(n,f)=1. \tag{31}$$

The EM algorithm is derived by setting the partial derivative of this Q-function by each parameter to be zero.

In the M-step, the variances $\nu_j(n, f)$, the spatial covariance matrix $\mathbf{R}_j(f)$ and the prior probability $\mu_j(f)$ are updated as

$$\nu_j(n,f)\leftarrow\frac{\mathbf{x}^H(n,f)\mathbf{R}_j^{-1}(f)\mathbf{x}(n,f)}{I}, \tag{32}$$

$$\mathbf{R}_j(f)\leftarrow\frac{\sum_n\frac{m_j(n,f)}{\nu_j(n,f)}\mathbf{x}(n,f)\mathbf{x}^H(n,f)}{\sum_n m_j(n,f)}, \tag{33}$$

$$\mu_j(f)\leftarrow\frac{\sum_n m_j(n,f)}{\sum_{n,j'}m_{j'}(n,f)}. \tag{34}$$

In E-step, the posterior probabilities are updated.

$$m_j(n,f)\leftarrow\frac{\mu_j(f)\mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f);\mathbf{0},\nu_j(n,f)\mathbf{R}_j(f)\right)}{\sum_{j'}\mu_{j'}(f)\mathcal{N}_{\mathrm{c}}\left(\mathbf{x}(n,f);\mathbf{0},\nu_{j'}(n,f)\mathbf{R}_{j'}(f)\right)}. \tag{35}$$

C. Source Separation

Here we formulate three source separation schemes using the estimated parameters.

1) Sparse soft mask using the posterior probability: The first scheme is to use the posterior probability $m_j(n, f)$ directly as the mask.

$$\mathbf{c}_j(n,f)=m_j(n,f)\mathbf{x}(n,f). \tag{36}$$

2) Sparse binary mask: The second scheme is using the binary mask $M_j(n, f)$ to regard the observation as the exclusive occurrence of one source with the maximum posterior probability $m_j(n, f)$, given by

$$M_j(n,f)=\begin{cases}1 & \text{if } \ j^*=\underset{j}{\operatorname{argmax}}\, m_j(n,f)\\ 0 & \text{otherwise}\end{cases}, \tag{37}$$

$$\mathbf{c}_j(n,f)=M_j(n,f)\mathbf{x}(n,f), \tag{38}$$

which corresponds to the MAP estimate of the source image $\mathbf{c}_j(n, f)$ to maximize $p(\mathbf{c}_1(n,f),...,\mathbf{c}_J(n,f)|\mathbf{x}(n,f),\mathbf{R}_1(f),...,\mathbf{R}_J(f),\nu_1(n,f),...,\nu_J(n,f))$.

3) Sparse Wiener filter: In the third scheme, we design a multichannel Wiener filter similar to the one in Duong's method, using the expectation of each variance $m_j(n, f)\nu_j(n, f)$;

$$\mathbf{R}_{\mathbf{x}}(n,f)=\sum_j m_j(n,f)\nu_j(n,f)\mathbf{R}_j(f), \tag{39}$$

$$\mathbf{R}_{\mathbf{c}_j}(n,f)=m_j(n,f)\nu_j(n,f)\mathbf{R}_j(f), \tag{40}$$

$$\hat{\mathbf{c}}_j(n,f)=\mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f). \tag{41}$$

In the next section, we conduct experiments to compare Duong's method without the assumption of sparseness to general binary mask MENUET and three methods with the assumption of sparseness formulated in this section.

V. Experiments

A. Experimental Condition

We compare the performances of MENUET the general binary masking method, Duong's method and the three source separation methods derived from the sparseness assumptions. We evaluated the source separation performance of six mixtures of three speeches chosen from four male and four female utterances as listed in Table I.

We used two-channel microphone array with the inter-element spacing of 2.15 cm. The distances between the sources and the microphone array are 1.0 m. The horizontal angles of the sources are $-40°$, $0°$ and $30°$. The sampling frequency is 16 kHz, and the frame size is 1024 samples. The EM algorithms are terminated after 50 iterations. Objective evaluation scores of separation performance are the following 4 distortion measures proposed in [12].

- SDR (Signal to Distortion Ratio): comprehensive distortion
- ISR (Source Image to Spatial distortion Ratio): linear distortion
- SIR (Source to Interference Ratio): distortion by the rest reduction of the other source
- SAR (Sources to Artifacts Ratio): non-linear distortion

The unit is dB, and the higher values show the better performance. We used the medhod in [8] to generate the initial values and to align the permutation for Duong's method and the formulated methods. The other experimental conditions are listed in Table II.

TABLE I
Combination of signal mixture

	$-40°$	$0°$	$30°$
1	Male1	Female2	Female3
2	Female1	Male2	Female4
3	Female2	Female4	Male3
4	Female1	Male1	Male4
5	Male4	Female2	Male2
6	Male1	Male3	Female3

TABLE II
Experimental condition

sources	SiSEC 2011
source direction	$-40°$, $0°$, $30°$
number of microphones	2
distance between the two microphones	2.15 cm
FFT size	1024 point
sampling frequency	16000 kHz
iteration of EM algorithm	50 times
performance measures	SDR, ISR, SIR, SAR

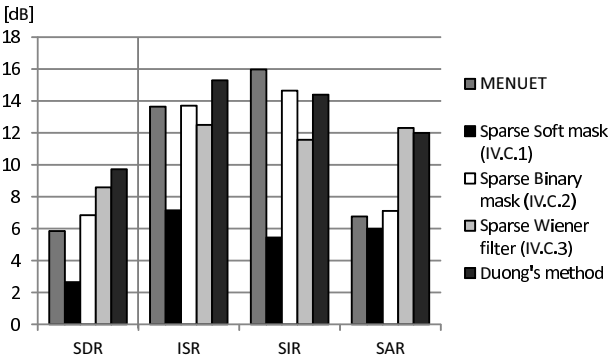

Fig. 1. Experimental result

B. Discussion

The experimental results are shown in Fig. 1. Duong's method marked good score in SIR, which indicates degree of separation, and the highest score in SDR, which measures the overall quality of the separated signal. Thus Duong's method is the best in this comparison. We discuss in detail in the following.

In order to verify the validity of the multivariate modeling of source image, we discuss the comparison between MENUET and the sparse binary mask. Since all the evaluation scores are similar, the multivariate model seems to have no remarkable side effect and its validity is confirmed.

In order to verify which source separation scheme is effective, we compare the three methods formulated by the sparse model. The sparse soft mask has the lowest scores. And while the sparse binary mask marks higher scores in ISR and SIR than the sparse Wiener filter, the sparse Wiener filter marked the higher scores in SDR and SAR. The assumption of the binary mask, which regards that in each time-frequency slot the estimated dominant source solo generates the observed signal, is effective for yielding high SIR, the source separation score. However, this assumption is not strictly satisfied, and the resultant error causes high distortion, as can be seen in the low SAR of the sparse binary mask. Because of the high distortion, SDR of the binary mask, which measures the overall quality of the separated signal, is not very high in spite of the high SIR. In contrast, the sparse Wiener filter has low distortion as can be seen in its high SAR. Thus SDR of the sparse Wiener filter is higher than that of the sparse binary mask even with the lower SIR, and the sparse Wiener filter has the better source separation performance than the sparse binary and soft maskings. Therefore we can conclude that Wiener filter performs the high-quality source separation with lower distortion than masking.

Moreover, in order to verify the effectiveness of source image superposition model not assuming the sparseness, we compare the performances of the formulated sparse Wiener filter and Duong's method. These two methods has similar SAR and causes similarly low amounts of distortion. However, Duong's method marks higher scores in SIR and ISR, and

as a result Duong's method has higher SDR than the sparse Wiener filter. Thus we can say that Duong's method has better source separation performance than the sparse Wiener filter, and the best one among those compared methods. As we discussed in Sect. III-C, Duong's method estimates the parameters of the multichannel Wiener as the optima of the assumed probabilistic model in two senses of expectation and MAP. However, the parameters of the sparse Wiener filter is estimated indirectly in its probabilistic model combining the estimations of soft mask and the covariance matrices, as shown in Sect. IV-C. Therefore, we confirmed the effectiveness of the parameter estimation by Duong's method without assuming sparseness both theoretically and experimentally.

As a result of the above comparisons, we can conclude that multichannel Wiener filter causes low distortion, and by effective estimation of the parameters employing the multivariate model without assuming sparseness, Duong's method performs high-quality and high-performance source separation.

VI. Conclusion

To investigate the factor of the low-distortion source separation by Duong's method, we analyzed the three characteristic attributes of Duong's method, i.e., source image model with multivariate normal distribution, separation by multichannel Wiener filter and source observation model of superimposition of source images without assuming sparseness among sources. For this analysis, we formulated three alternative BSS methods by sparse soft mask, sparse binary mask and sparse Wiener filter, and the performances are compared with the general binary masking method MENUET and Duong's method. All of these methods are derived from the same source image model of Duong's method and different observation model where the sources are not superimposed but a single source image appears in each time-frequency slot assuming source sparseness. As a result, effectiveness of all the above three attributes of Duong's method has been confirmed.

Acknowledgment

This work was supported by MEXT/JSPS KAKENHI Grant Number 23240023. We would like to thank Dr. Masato Togami of Central Research Laboratory, Hitachi, Ltd., who gave us fruitful suggestions about the probabilistic model of Duong's method.

References

[1] H. Sawada, S. Araki, and S. Makino, "Recent advances in audio source separation techniques," *J. IEICE,* Vol. 91, No. 4, pp. 292–296, 2008.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis,* Wiley, New York, 2001.

[3] O. Yilmaz, S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.,* vol. 52, No. 7, pages 1830–1847, 2004.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing,* no. 87, pp. 1833–1847, 2007.

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

[6] O. L. Frost. "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE,* vol. 60, no. 8, pp. 926–935, August 1972.

[7] J. Cermak, S. Araki, H. Sawada, S. Makino, "Blind speech separation by combining beamformers and a time frequency binary mask," *Proc. IWAENC,* pp. 145–148, 2006.

[8] K. Iso, S. Araki, and S. Makino, T. Nakatani, H. Sawada, T. Yamada, and A. Nakamura, "Blind source separation of mixed speech in a high reverberation environment," *Proc. HSCMA,* pp. 36–39, 2011.

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

[10] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio Speech Language Process.,* vol. 18, no. 7, pp. 1830–1840, 2010.

[11] M. Togami, "Statistical estimation theory considering time-varying nature of systems and source-probability distributions," *Ph. D. thesis,* the University of Tokyo, 2011.

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