Bayesian Unification of Sound Source Localization and Separation with Permutation Resolution

Takuma Otsuka†, Katsuhiko Ishiguro‡, Hiroshi Sawada‡, and Hiroshi G. Okuno†
†Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan.
{*ohtsuka, okuno*}*@kuis.kyoto-u.ac.jp*
‡NTT Communication Science Laboratories, NTT Corporation, Kyoto, 619-0237, Japan.
{*ishiguro.katsuhiko, sawada.hiroshi*}*@lab.ntt.co.jp*

Abstract

Sound source localization and separation with permutation resolution are essential for achieving a computational auditory scene analysis system that can extract useful information from a mixture of various sounds. Because existing methods cope separately with these problems despite their mutual dependence, the overall result with these approaches can be degraded by any failure in one of these components. This paper presents a unified Bayesian framework to solve these problems simultaneously where localization and separation are regarded as a clustering problem. Experimental results confirm that our method outperforms state-of-the-art methods in terms of the separation quality with various setups including practical reverberant environments.

1 Introduction

Computational auditory scene analysis (CASA) seeks for the intelligence capable of the analysis and comprehension of ambient or recorded auditory events in terms of their speech content or type of sound source (Rosenthal and Okuno 1998; Wang and Brown 2006). This technology contributes to a better understanding of auditory events as well as an agent capable of oral interaction with human beings (Nakadai et al. 2000). Because most environments contain multiple sounds, many CASA systems use multiple sensors, namely a microphone array, to extract useful information from the observed audio mixture (Benesty, Chen, and Huang 2008).

CASA systems may conform to the three-layer architecture shown in Figure 1, which is similar to the subsumption architecture (Brooks 1986) in that each layer deals with a specific goal and higher layers become more abstract; (1) a sensor arrangement and a microphone array configuration (Hulsebos, de Vries, and Bourdillat 2002), (2) the decomposition of the observed sound mixture into individual sources such as sound source localization and separation (Asano et al. 2001; Nakadai et al. 2010), and (3) an analysis of each source, e.g., speech recognition or speaker identification (McTear 2004; Yamada, Sugiyama, and Matsui 2010). This paper focuses on the 2nd layer because the decomposition quality generally affects the subsequent applicative processes such as speech recognition.

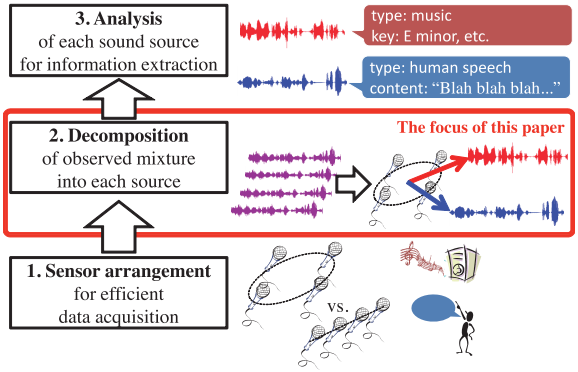

Figure 1: Three-layer architecture for CASA systems

We tackle sound source separation along with permutation resolution and localization as the decomposition problem. The goal of sound source separation is to retrieve sound sources from an observed sound mixture to facilitate the analysis of each source (Common and Jutten 2010; Lee, Kim, and Lee 2007; Sawada, Araki, and Makino 2011). Some separation methods require permutation resolution (Sawada et al. 2004); when an unsupervised separation is carried out in the time-frequency domain, separated signals of the same source should be aggregated from all frequency bins to resynthesize the original signal. Finally, the localization aims to estimate the arrival direction of each sound source (Asano and Asoh 2011). This information is useful for robots aware of ambient auditory events (Wang, Ivanov, and Aarabi 2003; Sasaki, Kagami, and Mizoguchi 2009), or displaying the activity in a meeting involving multiple people (Kubota et al. 2008).

While the problems of sound source localization, separation, and permutation resolution are mutually dependent, most existing methods deal with a specific part of these problems, and are combined to handle a compound problem. The metaphor of Liebig's barrel applies to the disadvantage of the combinational approach, namely, the overall quality is determined by the worst component. For example, a failure in the localization would degrade the separation, or an incorrect permutation resolution would result in a degraded audio signal. Our contribution consists of the formulation of a probabilistic generative model to solve these problems simultaneously in a unified Bayesian manner. The advantage of our method is

in that it both solves the multiple problems and improves the separation quality by incorporating a localization estimate.

2 Problem statement and issues

This section presents the problem dealt with in this paper, three issues, and the need for a unified framework. The decomposition problem is summarized as follows:

Input: Multichannel audio signal,
Outputs: The direction of arrival and separated signal of each sound source,
Assumptions: The microphone array configuration is known as steering vectors, the number of sources is given, and the sound source directions do not change over time.

A steering vector conveys the time difference of sound arrivals at each sensor given a certain direction of the sound source and a frequency bin. Our implementation uses steering vectors measured in an anechoic chamber. Thus, our method works independently of the environment. The number of sources is important especially in reverberant environments where automatic estimation is still a challenge. Some alleviation of these assumptions is discussed in Section 5.

The issues are threefold; (1) permutation resolution with frequency domain processing, (2) underdetermined conditions where the number of sources exceeds the number of sensors, and (3) distinguishable sound source selection.

Three issues First, most environments contain reverberation modeled as a convolutive mixture (Pedersen et al. 2007). Microphone array processing in the frequency domain copes with this situation. Although this strategy can handle the reverberation by converting the convolution into element-wise multiplications through a Fourier transform, the permutation problem (Sawada et al. 2004) is induced.

The permutation problem occurs when the separation is carried out independently of each frequency bin in an unsupervised manner, for example using independent component analysis (ICA) (Common and Jutten 2010). When we aggregate the spectrogram of a certain source, we must identify signals of the same sound source from all frequency bins. Independent vector analysis (IVA) (Lee, Kim, and Lee 2007; Ono 2011) avoids the permutation problem by maximizing the independence of constituent sound sources across all frequency bins at the same time.

Second, some linear mixture models including ICA and IVA assume that the number of sources N does not exceed the number of microphones M. In practice, however, N is not always guaranteed to be capped at M. The case where $N > M$ is called an underdetermined condition. While Sawada et al. (2011) cope with the underdetermined condition by the frequency-wise clustering of sound sources, their method requires permutation resolution after the clustering. This two-step strategy may degrade the overall separation quality.

The third issue is typically related to linear mixture models such as IVA (Lee, Kim, and Lee 2007; Ono 2011). In many cases, we have to select N distinct sound sources from among the M separated signals because in practice we often set $M > N$ to avoid underdetermined conditions. A failure in this selection step would degrade the separation itself as well as any subsequent analysis such as speech recognition.

The IVA algorithm simply decomposes an observed mixture consisting of M channels into M independent signals. In many cases, IVA is applied where $M > N$. Thus, we have to reduce the dimensionality M to N by employing, for example, principal component analysis as a preprocessing.

Towards a unified framework As explained in Section 1, the relationship between sound source localization, separation, and the permutation resolution obeys the Liebig's Law. For example, (Nakadai et al. 2010) achieves both localization and separation by solving each problem in order; after localizing each sound, each sound source is extracted by emphasizing signals coming from the estimated directions. This method is advantageous in that the separation avoids the permutation resolution because each source can be specified by its direction. Here, incorrect localization would result in degraded separation quality. The localization with IVA is another example: The correlation of the separated signals and steering vectors is investigated for the localization. Here, the localization process is sensitive to the preceding separation.

Two major solutions have been proposed for the permutation problem (Sawada et al. 2004). One involves synchronizing the change in the power envelope of a source over frequency bins. The other involves combining the separated signals estimated to come from the same direction. Both methods can degrade the outcome due to permutation errors by incorrect frequency-wise separation or localization.

Our method unifies all the problems into a Bayesian framework. The separation and localization problem is formulated as a clustering problem in the time-frequency domain. Permutation resolution employs both the above ideas: 1) We introduce a sound dominance proportion for each time frame to encourage each sound source to synchronously increase its power over frequency bins. 2) The direction is used to identify the separated signals at each frequency bin.

3 Method

Figure 2 outlines our method. First, the mixed signal to be observed is generated by adding sound sources as shown on the left in Fig. 2. A real-valued waveform in the time domain is converted into complex values in the time-frequency domain by a short-time Fourier transform (STFT). Then, a time-frequency mask (TF mask) is estimated for each source to retrieve it from the mixture.

Figure 2 shows power spectrograms on a linear scale to emphasize that the power is sparsely distributed in the time-frequency domain, that is, the power is nearly zero at most time-frequency points. Therefore, we can assume that only one sound source is dominant at each time-frequency point and that we are able to extract sound sources with TF masks.

The estimation of the TF masks is formulated as a clustering problem on the observed multi-channel signal in the time-frequency domain. Each time-frequency point stems from a certain source referred to as a class in the clustering context. Our method estimates the posterior probability to which class each time-frequency point belongs. Furthermore,

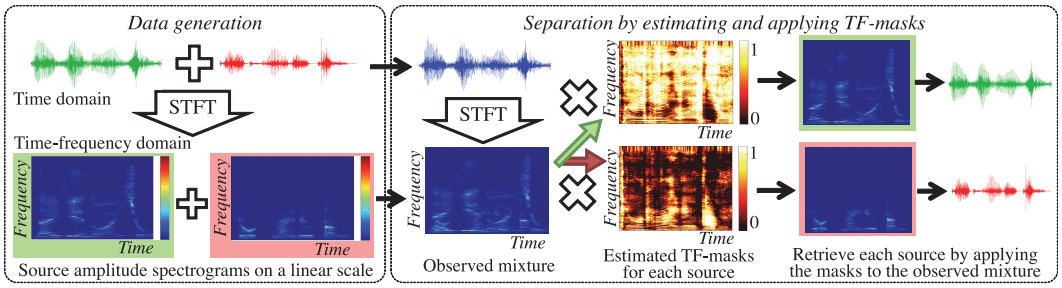

Figure 2: Illustration of mixture process and separation method based on time-frequency masking

our method handles more classes than the actual number of sound sources for the clustering to make our algorithm independent of the number of actual sound sources. We set parameters s.t. redundant classes shrink during the clustering and we obtain stable results regardless of the source number.

Our method resembles (Mandel, Ellis, and Jebara 2007) in that time-frequency points are clustered into each source to generate TF masks. While Mandel et al. use only the phase of complex values and 2 microphones, our method uses the complex values themselves and allows 2 or more microphones that produce better results.

Sections 3.1–3.3 explain the generative process and Section 3.4 presents the inference procedures. Table 1 shows the notations we use. A set of variables is denoted with a tilde without subscripts, e.g., $\tilde{\mathbf{x}} = \{\mathbf{x}_{tf} | 1 \leq t \leq T, 1 \leq f \leq F\}$.

Table 1: Notations

Symbol	Meaning
t	Time frame ranging from 1 to T
f	Frequency bin from 1 to F
k	Class index from 1 to K
d	Direction index from 1 to D
M	Number of microphones
N	Number of sound sources
$\mathbf{x}_{tf}$	Observed M-dimensional complex column vector
$\mathbf{z}_{tf}$	Class indicator at t and f
$\boldsymbol{\pi}_t$	Class proportion at time t
$\mathbf{w}_k$	Direction indicator for class k
$\boldsymbol{\varphi}$	Direction proportion for all classes
λ_{tfk}	Inverse power of class k at t and f
$\mathbf{H}_{fd}$	Inverse covariance of direction d at frequency f

3.1 Observation model with time-varying covariance matrices

We employ the covariance model (Duong, Vincent, and Gribonval 2010) for the likelihood function of the signal in the time-frequency domain; each sample follows a complex normal distribution with zero mean and time-varying covariance. Figure 3 shows a scatter plot of the two-channel observations of two sources in blue and red. These samples are generated as follows; let s_{tfk} and $\mathbf{q}_{fd}$ denote the signal of the kth class at time t and frequency f, and the steering vector from direction d where class k is located. Then, the signal is observed as $\mathbf{x}_{tf} = s_{tfk}\mathbf{q}_{fd}$, where the elements of $\mathbf{x}_{tf}$ are the observation of each microphone. The covariance is

$$\mathbb{E}[\mathbf{x}_{tf}\mathbf{x}_{tf}^H] = \mathbb{E}[|s_{tfk}|^2 \mathbf{q}_{fd}\mathbf{q}_{fd}^H], \quad (1)$$

where $\cdot^H$ means a Hermitian transpose.

As shown in Figure 3, the covariance matrix of each source has an eigenvector with a salient eigenvalue. This vector corresponds to the steering vector associated with the direction in which the source is located. That is, the clustering of each sample corresponds to the separation of sound sources, and the investigation of the eigenvectors of the clustered covariances means the localization of sources.

The covariance is factorized into a power term and a steering matrix. While the power of the signal $|s_{tfk}|^2$ is time-varying in Eq. (1), the steering term $\mathbf{q}_{fd}\mathbf{q}_{fd}^H$ is fixed over time since we assume steady sources. Because we can assume $s_{t,f,k}$ and $\mathbf{q}_{fd}$ are independent, we introduce an inverse power $\lambda_{tfk} \approx |s_{tfk}|^{-2}$ and an inverse steering matrix $\mathbf{H}_{fd} \approx (\mathbf{q}_{fd}\mathbf{q}_{fd}^H + \epsilon\mathbf{I}_M)^{-1}$, where $\mathbf{I}_M$ is the $M \times M$ identity matrix. The likelihood distribution is

$$p(\tilde{\mathbf{x}}|\tilde{\mathbf{z}}, \tilde{\mathbf{w}}, \tilde{\lambda}, \tilde{\mathbf{H}}) = \prod_{tfkd} \mathcal{N}_{\mathbb{C}}(\mathbf{x}_t f|\mathbf{0}, (\lambda_{tfk}\mathbf{H}_{fd})^{-1})^{z_{tfk}w_{kd}}, \quad (2)$$

where $\mathcal{N}_{\mathbb{C}}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) = \frac{|\boldsymbol{\Lambda}|}{(2\pi)^M} \exp\left(-\mathbf{x}^H\boldsymbol{\Lambda}\mathbf{x}\right)$ is the probability density function (pdf) of the complex normal distribution (van den Bos 1995) with a mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$. $\mathbf{z}_{tf} = [z_{tf1}, ..., z_{tfK}]$ and $\mathbf{w}_k = [w_{k1}, ..., w_{kD}]$ indicates the class of $\mathbf{x}_{tf}$ and the direction of class k, respectively. In this vector representation, one of the elements equals 1 and the others are 0; if the class is k' at t and f, $z_{tfk'} = 1$ and $z_{tfk} = 0$ for any other k. By placing these binary variables in the exponential part of the likelihood function and calculating the product over all possible k and d, Eq. (2) provides the likelihood given the class and its direction. $|\boldsymbol{\Lambda}|$ is the determinant of the matrix $\boldsymbol{\Lambda}$.

We adopt conjugate priors for parameters λ_{tfk} and $\mathbf{H}_{fd}$:

$$p(\tilde{\lambda}) = \prod_{tfk} \mathcal{G}(\lambda_{tfk}|a_0, b_{tf}), \quad (3)$$

$$p(\tilde{\mathbf{H}}) = \prod_{fd} \mathcal{W}_{\mathbb{C}}(\mathbf{H}_{fd}|\nu_0, \mathbf{G}_{fd}), \quad (4)$$

where $\mathcal{G}(\lambda|a, b) \propto \lambda^{a-1}e^{-b\lambda}$ denotes the pdf of a gamma distribution with a shape a and inverse scale b, and $\mathcal{W}_{\mathbb{C}}(\mathbf{H}|\nu, \mathbf{G}) = \frac{|\mathbf{H}|^{\nu-M}\exp\{-\mathrm{tr}(\mathbf{H}\mathbf{G}^{-1})\}}{|\mathbf{G}|^{\nu}\pi^{M(M-1)/2}\prod_{i=0}^{M-1}\Gamma(\nu-i)}$ is the pdf of a complex Wishart distribution (Conradsen et al. 2003). $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$ and $\Gamma(x)$ is the gamma function.

Hyperparameters are set as: $a_0 = 1$, $b_{tf} = \mathbf{x}_{tf}^H\mathbf{x}_{tf}/M$, $\nu_0 = M$, $\mathbf{G}_{fd} = (\mathbf{q}_{fd}\mathbf{q}_{fd}^H + \epsilon\mathbf{I}_M)^{-1}$. The gamma parameter

b_{tf} reflects the power of the observation and the Wishart parameter $\mathbf{G}_{fd}$ is generated from the given steering vectors $\mathbf{q}_{fd}$ where $\mathbf{q}_{fd}$ is normalized s.t. $\mathbf{q}_{fd}^H \mathbf{q}_{fd} = 1$, and $\epsilon = 0.001$ to allow the inverse operation.

3.2 Permutation resolution based on LDA

The bottom right image in Figure 4 shows how dominant each source is at time frames. As the figure shows, the red source is dominant in some time frames whereas the blue source is dominant in others. We can expect to resolve the permutation by preferring one or several classes for each time frame in a way similar to that used in (Sawada, Araki, and Makino 2011) to seek the synchronization of the sound dominance over frequency bins.

Here, we use a topic model called latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) to introduce the proportion of classes. LDA infers the topic of documents containing many words from a document set by assigning each word to a certain topic. We regard the topic as a sound source, the document as a time frame, and the words as frequency bins.

Let $\boldsymbol{\pi}_t$ denote the class proportion at time t hereafter. The class indicator variable $\mathbf{z}_{tf}$ in Eq. (2) determines which class $\mathbf{x}_{tf}$ belongs to in accordance with $\boldsymbol{\pi}_t$ as:

$$p(\tilde{\mathbf{z}}|\tilde{\boldsymbol{\pi}}) = \prod_{tfk} \pi_{tk}^{z_{tfk}}, \tag{5}$$

where $\boldsymbol{\pi}_t$ follows a conjugate prior Dirichlet distribution:

$$\begin{aligned} p(\tilde{\boldsymbol{\pi}}|\boldsymbol{\beta}) &= \prod_t \mathcal{D}(\boldsymbol{\pi}_t|\alpha\boldsymbol{\beta}) \\ &= \prod_t \frac{\Gamma(\alpha\beta_\cdot)}{\prod_k \Gamma(\alpha\beta_k)} \prod_k \pi_{tk}^{\alpha\beta_k - 1}, \end{aligned} \tag{6}$$

where the subscript $\cdot$ denotes the summation over the specified index, i.e., $\beta_\cdot = \sum_k \beta_k$.

The global class proportion $\boldsymbol{\beta} = [\beta_1, ..., \beta_K]$ is made asymmetric to encourage the shrinkage of redundant classes (Wallach, Mimno, and McCallum 2009). The construction is similar to the stick-breaking process (Sethuraman 1994). In contrast to the sampling procedure in the ordinary stick-breaking process, we use the expectation of stick segments to avoid the numerical problem of any β_k being zero:

$$\beta'_k = \mathbb{E}[\mathcal{B}(\beta'|1,\gamma)] = 1/(1+\gamma), \tag{7}$$

$$\beta_1 = \beta'_1, \quad \beta_k = \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l), \text{ for } k = 2 \ldots K, \tag{8}$$

where $\mathbb{E}[\mathcal{B}(x|\alpha,\beta)]$ is the expectation of a beta distribution with parameters α and β. These parameters are given as: $\gamma = 0.2$ and $\alpha = 0.2$.

3.3 Latent variable for localization

A discrete variable $\mathbf{w}_k = [w_{k1}, ..., w_{kD}]$ that indicates the direction d of each class k is introduced to localize each sound source and to solve the permutation problem by employing the sound location. The directions of the sound sources are discrete in this model to simplify the inference process.

The indicator $\mathbf{w}_k$ is dependent on the direction proportion $\boldsymbol{\varphi}$ that follows a Dirichlet distribution:

$$p(\tilde{\mathbf{w}}|\boldsymbol{\varphi}) = \prod_{kd} \varphi_d^{w_{kd}}, \tag{9}$$

$$p(\boldsymbol{\varphi}) = \mathcal{D}(\boldsymbol{\varphi}|\kappa\mathbf{1}_D) = \frac{\Gamma(D\kappa)}{\prod_d \Gamma(\kappa)} \prod_d \varphi_d^{\kappa-1}, \tag{10}$$

where $\mathbf{1}_D$ is a D-dimensional vector whose elements are all 1. Note that we use a symmetric Dirichlet distribution with concentration parameter κ because we have no prior knowledge about the spatial position of the sound sources. The hyperparameter κ is set uninformative: $\kappa = 1$.

3.4 Inference

Figure 5 depicts the probabilistic dependency; the double-circled $\mathbf{x}_{tf}$ is the observation, the circled symbols are latent probability variables, and the plain symbols are fixed values.

The posterior distribution over all latent variables given the observation is estimated by the variational Bayesian inference (Attias 2000), as derived in detail in (Bishop 2006). The posterior is approximated by factorized distributions q:

$$p(\tilde{\mathbf{z}}, \tilde{\mathbf{w}}, \tilde{\lambda}, \tilde{\mathbf{H}}, \tilde{\boldsymbol{\pi}}, \boldsymbol{\varphi}|\tilde{\mathbf{x}}) \approx q(\tilde{\mathbf{z}})q(\tilde{\mathbf{w}})q(\tilde{\lambda})q(\tilde{\mathbf{H}})q(\tilde{\boldsymbol{\pi}})q(\boldsymbol{\varphi}). \tag{11}$$

During the inference process, we update one of the factorized distributions in Eq. (11) while fixing the other distributions s.t. the following objective function is maximized:

$$\begin{aligned} \mathcal{L}(q) = &\mathbb{E}_q[\log p(\tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}}, \tilde{\lambda}, \tilde{\mathbf{H}}, \tilde{\boldsymbol{\pi}}, \boldsymbol{\varphi})] \\ &- \mathbb{E}_q[\log q(\tilde{\mathbf{z}})q(\tilde{\mathbf{w}})q(\tilde{\lambda})q(\tilde{\mathbf{H}})q(\tilde{\boldsymbol{\pi}})q(\boldsymbol{\varphi})], \end{aligned} \tag{12}$$

where $\mathbb{E}_q[\cdot]$ means the expectation over distribution q in Eq. (11). Note that the joint distribution p of all variables in Eq. (12) is the product of Eqs. (2–6, 9, 10).

The choice of conjugate priors enables factorized posteriors to conform to the same distributions as the priors:

$$\begin{aligned} q(\tilde{\mathbf{z}}) &= \prod_{tfk} \xi_{tfk}^{z_{tfk}}, \quad q(\tilde{\mathbf{w}}) = \prod_{kd} \eta_{kd}^{w_{kd}}, \\ q(\tilde{\lambda}) &= \prod_{tfk} \mathcal{G}(\lambda_{tfk}|\hat{a}_{tfk}, \hat{b}_{tfk}), \\ q(\tilde{\mathbf{H}}) &= \prod_{fd} \mathcal{W}_{\mathbb{C}}(\mathbf{H}_{fd}|\hat{\nu}_{fd}, \hat{\mathbf{G}}_{fd}), \\ q(\tilde{\boldsymbol{\pi}}) &= \prod_t \mathcal{D}(\boldsymbol{\pi}_t|\hat{\boldsymbol{\beta}}_t), \quad q(\boldsymbol{\varphi}) = \mathcal{D}(\boldsymbol{\varphi}|\hat{\boldsymbol{\kappa}}). \end{aligned} \tag{13}$$

The parameters in Eqs. (13) are updated as follows:

$$\begin{aligned} \log \xi_{tfk} &= \psi(\hat{\beta}_{tk}) - \psi(\hat{\beta}_{t\cdot}) + M\mathbb{E}_q[\log \lambda_{tfk}] \\ &\quad + \sum_d \eta_{kd} \left\{ \mathbb{E}_q[\log |\mathbf{H}_{fd}| - \lambda_{tfk} \mathbf{x}_{tf}^H \mathbf{H}_{fd} \mathbf{x}_{tf}] \right\} + C, \\ \log \eta_{kd} &= \psi(\hat{\kappa}_d) - \psi(\hat{\kappa}_\cdot) \\ &+ \sum_{tf} \xi_{tfk} \left\{ M\mathbb{E}_q[\log \lambda_{tfk} + \log |\mathbf{H}_{fd}| - \lambda_{tfk} \mathbf{x}_{tf}^H \mathbf{H}_{fd} \mathbf{x}_{tf}] \right\} + C, \end{aligned} \tag{14}$$

$$\begin{aligned} &\hat{\beta}_{tk} = \beta_k + \xi_{t\cdot k}, \quad \hat{\kappa}_d = \kappa + \eta_{\cdot d}, \\ &\hat{a}_{tfk} = a_0 + M\xi_{tfk}, \quad \hat{b}_{tfk} = b_{tf} + \xi_{tfk} \sum_d \eta_{kd} \hat{\nu}_{fd} \mathbf{x}_{tf}^H \hat{\mathbf{G}}_{fd} \mathbf{x}_{tf}, \\ &\hat{\nu}_{fd} = \nu_0 + \sum_{tk} \xi_{tfk} \eta_{kd}, \quad \hat{\mathbf{G}}_{fd}^{-1} = \mathbf{G}_{fd}^{-1} + \sum_{tk} \xi_{tfk} \eta_{kd} \frac{\hat{a}_{tfk}}{\hat{b}_{tfk}} \mathbf{x}_{tf} \mathbf{x}_{tf}^H. \end{aligned} \tag{15}$$

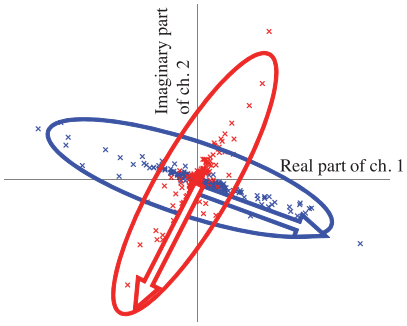

Figure 3: Plot of complex-valued multi-channel signals at 3000 (Hz). The colors represent respective sound sources.

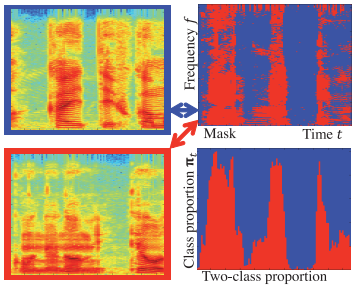

Figure 4: TF mask for two sources (top right) and their proportion for each time frame (bottom right). Left: log-scale power spectrograms of the two sources in blue and red.

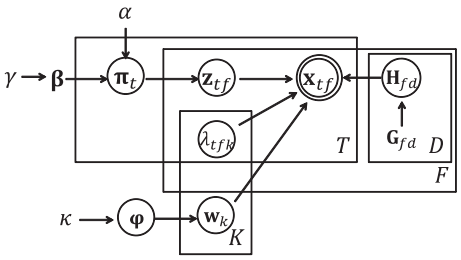

Figure 5: Graphical representation of our model.

Eq. (14) has constant terms C such that ξ_{tfk} and η_{kd} are normalized as $\xi_{tf\cdot} = 1$ and $\eta_{k\cdot} = 1$. The expectations over the gamma and complex Wishart distributions are given as:

$$\mathbb{E}_q[\log \lambda_{tfk}] = \psi(\hat{a}_{tfk}) - \log \hat{b}_{tfk},$$
$$\mathbb{E}_q[\log |\mathbf{H}_{fd}|] = \sum_{i=0}^{M-1} \psi(\hat{\nu}_{fd} - i) + \log |\hat{\mathbf{G}}_{fd}|,$$
$$\mathbb{E}_q[\lambda_{tfk} \mathbf{x}_{tf}^H \mathbf{H}_{fd} \mathbf{x}_{tf}] = \frac{\hat{a}_{tfk}}{\hat{b}_{tfk}} \mathbf{x}_{tf}^H \hat{\nu}_{fd} \hat{\mathbf{G}}_{fd} \mathbf{x}_{tf}.$$

Note that these expectations are over $q(\boldsymbol{\lambda})$, $q(\mathbf{H})$ or both in Eq. (13). $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function.

The updates in Eqs. (14–15) are iterated in turn until the objective function in Eq. (12) converges. During the iteration, the class index k is sorted by the total weights calculated by $\xi_{\cdot\cdot k}$ to accelerate the shrinkage of redundant classes (Kurihara, Welling, and Teh 2007).

Finally, the spectrogram of the nth sound source $\hat{\mathbf{x}}_{tf}^n$ is extracted as follows:

$$\hat{\mathbf{x}}_{tf}^n = \frac{\xi_{tfn}}{\sum_{k'=1}^{N} \xi_{tfk'}} \mathbf{x}_{tf}. \tag{16}$$

The weights ξ_{tfk} are normalized within a given number of sources N to obtain better separation quality. The direction of the nth sound $\hat{d}_n$ is obtained as

$$\hat{d}_n = \operatorname*{argmax}_{d'} \eta_{nd'}. \tag{17}$$

Initialization The inference begins by setting η_{kd} and ξ_{tfk}. First, η_{kd} is initialized s.t. each class takes on an equal range of directions. Then, ξ_{tfk} is set in accordance with the correlation between $\mathbf{x}_{tf}$ and designated directions. The equations are:

$$\eta_{kd} \propto \begin{cases} 1 & (k-1)D/K \le d < kD/K, \\ 0 & \text{otherwise}, \end{cases} \tag{18}$$

$$\xi_{tfk} \propto \exp\left\{ -\mathbf{x}_{tf}^H \sum_d (\eta_{kd} \mathbf{G}_{fd}) \mathbf{x}_{tf} \right\} \tag{19}$$

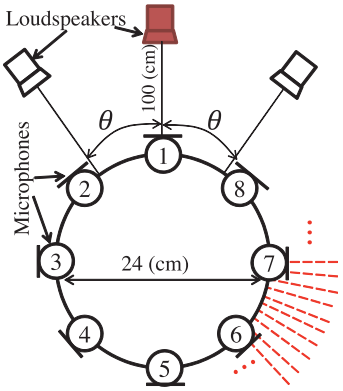

Figure 6: Experimental setup for the impulse response measurement shown in red dotted lines and simulated convolutive mixtures.

Note that a large K ensures that each class k contains at most one sound source. During the inference, classes lose their weight when they do not cover any direction in which a sound source actually exists, and classes associated with directions in the presence of sound sources increase in weight.

4 Experimental Results

This section presents localization and separation results obtained with simulated convolutive mixture signals. The experiments explore the way in which our method is influenced by the number of microphones and speakers, the amount of reverberation, and the interval between speakers. Our method is compared with state-of-the-art sound separation methods; IVA (Ono 2011) when $M \ge N$ and TF-masking with permutation resolution referred to as TF-Perm. (Sawada, Araki, and Makino 2011) when $M < N$.

4.1 Experimental setup

Figure 6 shows a circular microphone array and the location of the speakers. Impulse responses are measured with a 5° resolution around the array, namely $D = 72$, as drawn with

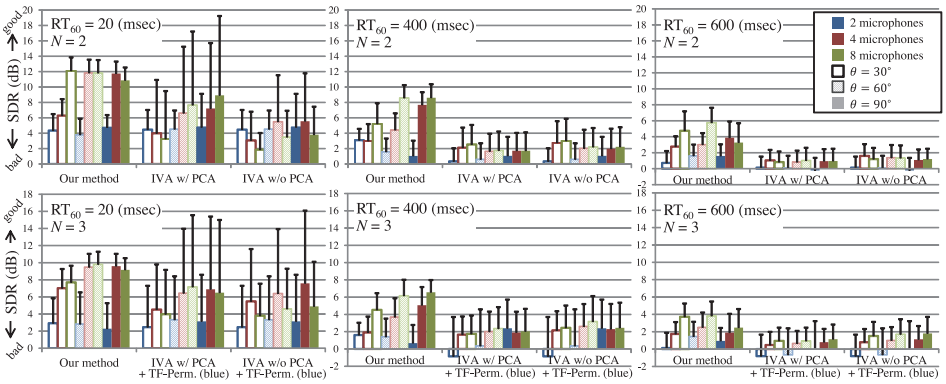

Figure 8: Separation scores in signal to distortion ratio (SDR). Larger values mean better separation results. The bars are the mean values and the segments are the standard deviations. Top: results with 2 sources. Bottom: results with 3 sources.

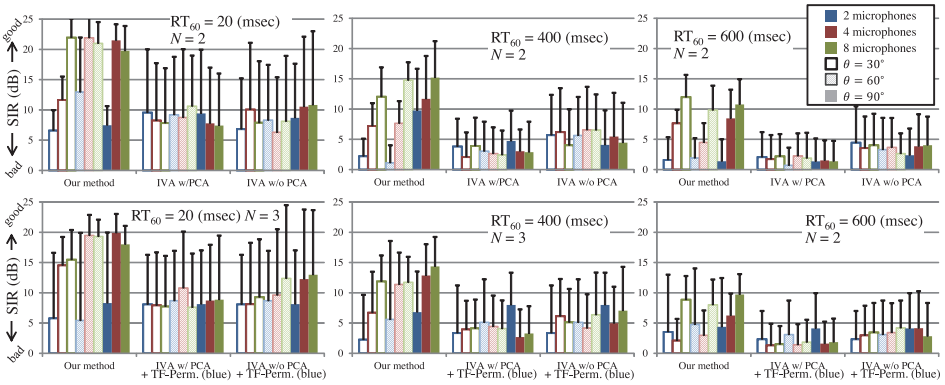

Figure 9: Signal to interference ratio (SIR). Larger values mean better quality. The bars are the mean values and the segments are the standard deviations. Top: results with 2 sources. Bottom: results with 3 sources.

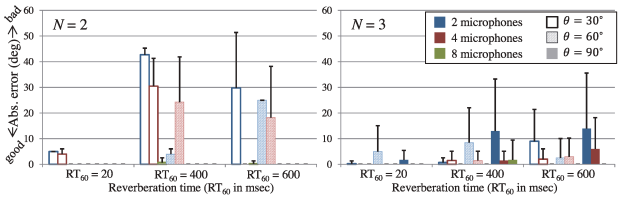

Figure 7: Absolute errors in degree in localization results. Smaller values mean better localization results. The bars are the mean values and the segments are the standard deviations. Left: results with 2 sources. Right: results with 3 sources.

dotted red lines in three environments where the reverberation times are $\mathrm{RT}_{60} = 20, 400, 600$ (msec), measured in an anechoic chamber, and two lecture rooms, respectively.

As depicted in Fig. 6, eight microphones are embedded with their channel number. The number of microphones M is 2, 4, or 8; channels 1 and 5 are used when $M = 2$, channels 1, 3, 5, and 7 when $M = 4$, and all channels when $M = 8$. Two or three sound sources are placed 100 (cm) from the array at an interval $\theta = 30, 60, 90°$. When two sources are present, the central source, shown in red in Fig. 6, is omitted. Therefore the interval becomes $2 \times \theta$. Under all conditions, the clustering is carried out with $K = 12$.

For each condition, 20 convolutive mixtures are generated from JNAS phonetically-balanced Japanese utterances. The speakers on the two sides in Fig. 6 are male and the center speaker is female. The audio signals are sampled at 16000 (Hz) and a short-time Fourier transform is carried out with a 512 (pt) hanning window and a 128 (pt) shift size. Steering vectors $\mathbf{q}$ are generated from a Fourier transform of the first 512 points of the anechoic impulsive responses.

Sound sources are selected as follows: Our method chooses the N most weighted classes. TF-Perm. assumes N sources to generate the TF masks. IVA has two strategies; (1) M-dimensional samples are preprocessed into the N-dimension by principal component analysis (IVA w/ PCA), and (2) N-first sources are chosen from M separated signals after being sorted in terms of signal power (IVA w/o PCA).

4.2 Results

Figure 7 shows the absolute localization errors with our method. The bars are the mean errors for 20 utterances under each condition and the segments are their standard deviation. The bar color represents the number of microphones, and the pattern shows the source intervals.

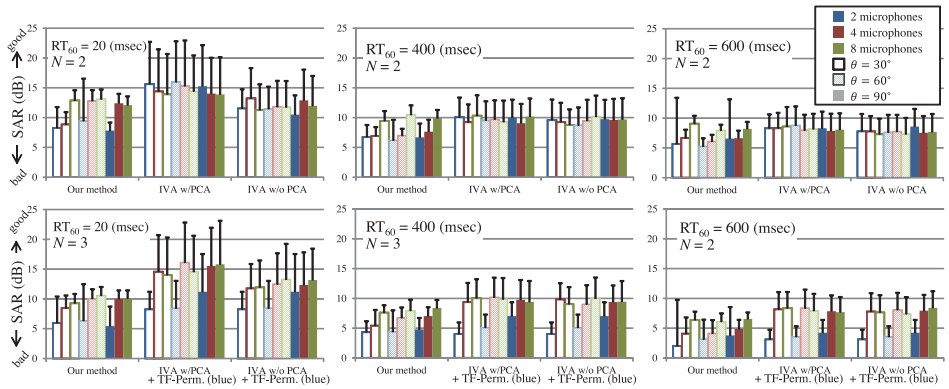

Figure 10: Signal to artifacts ratio (SAR). Larger values mean better quality. The bars are the mean values and the segments are the standard deviations. Top: results with 2 sources. Bottom: results with 3 sources.

The errors are suppressed with low reverberation $\mathrm{RT}_{60} = 20$ (msec) and with a larger number of microphones. In reverberant environments where $\mathrm{RT}_{60} = 400, 600$ (msec), localization errors are caused by the mismatched anechoic steering vectors and "*ghost*" sources reflected on walls.

The separation results are evaluated in terms of the signal to distortion ratio (SDR), signal to interference ratio (SIR), and signal to artifacts ratio (SAR) (Vincent, Gribonval, and Févotte 2006). SDR measures the overall retrieval quality of sound sources from their mixture while SIR measures how much the interfering other sources are removed and SAR shows the effect of artifacts caused by the separation process. Figure 8 shows mean SDR scores obtained with our method and competing methods; IVA w/ and w/o PCA when $M \geq N$, and TF-Perm. when $M < N$.

The mean SDR of our method is superior to that of competing methods whereas their standard deviation is larger. This is because the competing methods tend to extract one dominant source from the mixture while our method extracts all the sources with nearly equal quality. This result suggests that our sound selection by class weights is superior to the power-based selection or PCA-based preprocessing of IVA. In particular, when $\mathrm{RT}_{60} = 400, 600$ (msec), the competing techniques fail to extract distinct sources that are confused with reflected echoes. Furthermore, regardless of the relationship between M and N, our method outperforms the competing methods that switch between IVA and TF-Perm.

Similarly to the localization results, the separation quality is degraded by the smaller number of microphones as well as the reverberation. Reverberation hinders the shrinkage of redundant classes by adding weights to reflected sounds. A possible way of alleviating this problem is to incorporate dereverberation techniques (Yoshioka et al. 2011).

Figures 9 and 10 show the SIR and SAR scores, respectively. We can observe two tendencies in comparison between time-frequency clustering methods (our method and TF-Perm.) and the linear separation-based method (IVA). Time-frequency clustering methods tend to produce larger SIR and smaller SAR scores because explicit selection of sound sources at each time-frequency point can reduce the interference but cause the artifacts. Linear separation-based methods such as IVA can separate sound sources with less artifacts while some interference may remain.

Although our method requires steering vectors, these rely only on the form of the microphone array rather than on specific environments. The results show that the anechoic steering vectors are robust, especially for the separation task.

5 Discussion and Future work

The experiments used simulated convolutive mixtures of human speech. Future work includes an evaluation with actual recordings and other types of sounds such as music signals.

Our model is currently a finite-mixture model where we need to determine the number of classes K. An important extension of our method is the nonparametric Bayesian model where K is infinitely large. By using this model, we may be able to estimate the number of sources solely from the observation along with dereverberation techniques.

While a nonparametric model called infinite independent component analysis (Knowles and Ghahramani 2007) is applied to sound source separation with a microphone array, the model is limited to the time domain (Knowles 2007), which is extremely vulnerable to reverberation. The extension to time-frequency processing is necessary for the robustness against reverberation. We can expect the infinite extension of our method by incorporating (Teh et al. 2006; Wang, Paisley, and Blei 2011).

For moving sound sources, one possible extension is to make the direction variable $\mathbf{w}_k$ into a time-series sequence, e.g., $\mathbf{w}_{tk}$. We can naturally model the time-series data by introducing a hidden Markov model (MacKay 1997).

6 Conclusion

This paper presented a solution to sound source localization and separation with permutation resolution using a microphone array that is essential to CASA systems. Because the problems are mutually dependent, the compound problems are unified as a Bayesian clustering method. Experimental results confirmed that our method outperforms state-of-the-art separation methods under various conditions.

References

Asano, F., and Asoh, H. 2011. Joint Estimation of Sound Source Location and Noise Covariance in Spatially Colored Noise. In *Proc. of 19th European Signal Processing Conference*, 2009–2013.

Asano, F.; Goto, M.; Itou, K.; and Asoh, H. 2001. Real-time Sound Source Localization and Separation System and Its Application to Automatic Speech Recognition. In *Proc. of Eurospeech2001*, 1013–1016.

Attias, H. 2000. A Variational Bayesian Framework for Graphical Models. In *Advances in Neural Information Processing Systems 12*, 209–215.

Benesty, J.; Chen, J.; and Huang, Y. 2008. *Microphone Array Signal Processing*. Springer Topics in Signal Processing. Springer.

Bishop, C. M. 2006. *Pattern Recognition And Machine Learning*. Springer-Verlag. chapter 10: Approximate Inference.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

Brooks, R. 1986. A Robust Layered Control System for A Mobile Robot. *IEEE Journal of Robotics and Automation* 2(1):14–23.

Common, P., and Jutten, C., eds. 2010. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press.

Conradsen, K.; Nielsen, A. A.; Schou, J.; and Skiriver, H. 2003. A Test Statistic in the Complex Wishart Distribution and Its Application to Change Detection in Polarimetric SAR Data. *IEEE Trans. on Geoscience and Remote Sensing* 41(1):4–19.

Duong, N. Q. K.; Vincent, E.; and Gribonval, R. 2010. Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model. *IEEE Trans. on Audio, Speech, and Language Processing* 18(7):1830–1840.

Hulsebos, E.; de Vries, D.; and Bourdillat, E. 2002. Improved Microphone Array Configurations for Auralization of Sound Fields by Wave-Field Synthesis. *Journal of Audio Engineering Society* 50(10):779–790.

Knowles, D., and Ghahramani, Z. 2007. Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. In *Proc. of International Conference on Independent Component Analysis and Signal Separation*, 381–388.

Knowles, D. 2007. Infinite Independent Component Analysis. Technical report, MEng Information Engineering, Cambrige University.

Kubota, Y.; Yoshida, M.; Komatani, K.; Ogata, T.; and Okuno, H. G. 2008. Design and Implementation of 3D Auditory Scene Visualizer towards Auditory Awareness with Face Tracking. In *Proc. of IEEE International Symboposium on Multimedia (ISM-2008)*, 468–476.

Kurihara, K.; Welling, M.; and Teh, Y. W. 2007. Collapsed Variational Dirichlet Process Mixture Models. In *Proc. of International Joint Conferences on Artificial Intelligence*.

Lee, I.; Kim, T.; and Lee, T.-W. 2007. Fast Fixed-point Independent Vector Analysis Algorithms for Convolutive Blind Source Separation. *Signal Processing* 87(8):1859–1871.

MacKay, D. 1997. Ensemble Learning for Hidden Markov Models. Technical report, Department of Physics, Cambridge University.

Mandel, M. I.; Ellis, D. P. W.; and Jebara, T. 2007. An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments. *Advances in Neural Information Processing Systems* 19.

McTear, M. 2004. *Spoken Dialogue Technology*. London: Springer Verlag.

Nakadai, K.; Lourens, T.; Okuno, H. G.; and Kitano, H. 2000. Active Audition for Humanoid. In *Proc. of 17th National Conference on Artificial Intelligence*, 832–839.

Nakadai, K.; Takahashi, T.; Okuno, H. G.; Nakajima, H.; Yuji, H.; and Tsujino, H. 2010. Design and Implementation of Robot Audition System "HARK". *Advanced Robotics* 24(5–6):739–761.

Ono, N. 2011. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 189–192.

Pedersen, M. S.; Larsen, J.; Kjems, U.; and Parra, L. C. 2007. A Survey of Convolutive Blind Source Separation Methods. In Benesty, J.; Sondhi, M. M.; and Huang, Y., eds., *Springer Handbook of Speech Processing*. Springer Press.

Rosenthal, D. F., and Okuno, H. G. 1998. *Computational Auditory Scene Analysis*. New Jersey: Lawrence Erlbaum.

Sasaki, Y.; Kagami, S.; and Mizoguchi, H. 2009. Online Short-Term Multiple Sound Source Mapping for a Mobile Robot by Robust Motion Triangulation. *Advanced Robotics* 23(1–2):145–164.

Sawada, H.; Araki, S.; and Makino, S. 2011. Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment. *IEEE Trans. on Audio, Speech, and Language Processing* 19(3):516–527.

Sawada, H.; Mukai, R.; Araki, S.; and Makino, S. 2004. A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation. *IEEE Trans. on Audio, Speech, and Language Processing* 12(5):530–538.

Sethuraman, J. 1994. A Constructive Definition of Dirichlet Priors. *Statistica Sinica* 4:639–650.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* 101(476):1566–1581.

van den Bos, A. 1995. The Multivariate Complex Normal Distribution–A Generalization. *IEEE Trans. on Information Theory* 41(2):537–539.

Vincent, E.; Gribonval, R.; and Févotte, C. 2006. Performance Measurement in Blind Audio Source Separation. *IEEE Trans. on Audio, Speech, and Language Processing* 14(4):1462–1469.

Wallach, H.; Mimno, D.; and McCallum, A. 2009. Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems* 22:1973–1981.

Wang, D., and Brown, G. J. 2006. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York: Wiley-IEEE Press.

Wang, Q. H.; Ivanov, T.; and Aarabi, P. 2003. Acoustic Robot Navigation using Distributed Microphone Arrays. *Information Fusion* 5(2):131–140.

Wang, C.; Paisley, J.; and Blei, D. 2011. Online variational inference for the hierarchical Dirichlet process. In *Proc. Artificial Intelligence and Statistics*.

Yamada, M.; Sugiyama, M.; and Matsui, T. 2010. Semi-Supervised Speaker Identification Under Covariate Shift. *Signal Processing* 90(8):2353–2361.

Yoshioka, T.; Nakatani, T.; Miyoshi, M.; and Okuno, H. G. 2011. Blind Separation and Dereverberation of Speech Mixtures by Joint Optimization. *IEEE Trans. on Audio, Speech, and Language Processing* 19(1):69–84.