BLIND SEPARATION OF INFINITELY MANY SPARSE SOURCES

Hirokazu Kameoka[1,2], *Misa Sato*[3], *Takuma Ono*[1], *Nobutaka Ono*[4], *Shigeki Sagayama*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo
[2]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
[3]School of Engineering, The University of Tokyo
[4]Principles of Informatics Research Revision, National Institute of Informatics

ABSTRACT

This paper deals with the problem of underdetermined blind source separation (BSS) where the number of sources is unknown. We propose a BSS approach that simultaneously estimates the number of sources, separates the sources based on the sparseness of speech, and performs permutation alignment. We confirmed experimentally that reasonably good separation was obtained with the present method without specifying the number of sources.

Index Terms— Underdetermined blind source separation, sparseness, Dirichlet process, variational inference

1. INTRODUCTION

Blind Source Separation (BSS) is a technique for separating out individual source signals from microphone inputs when the transfer characteristics between sources and microphones are unknown. BSS is potentially useful for the development of such applications as hands-free teleconference systems and automatic meeting transcription systems. In meeting situations, for instance, it is likely that the number of participants (speakers) will change during the meeting or a loud, unexpected noise such as a door slamming will occur in the room. Thus, it is often difficult to pre-specify the exact number of all possible sources present in real environments. Many conventional BSS algorithms are designed to use the number of sources as the input when performing separation, and most of these algorithms do not work well if the assumed and actual numbers of sources are not the same. This paper presents a novel BSS algorithm that allows the number of sources to be inferred along with the separation.

To estimate the unknown mixing matrix and source signals solely from observed signals, we must make some assumptions about the sources, and formulate an appropriate optimization problem based on criteria designed according to those assumptions. For example, if the observed signals outnumber the sources, we can employ independent component analysis (ICA) [1], which estimates the separation matrix (the inverse of the mixing matrix) such that the independence of the source estimates is maximized. However, to achieve a BSS algorithm that works without assuming the number of sources, we shall always consider an underdetermined case where there are fewer observations than sources. In an underdetermined case, there are an infinite number of solutions for source signals even if the mixing process is known. The independence assumption is too weak to allow us to determine a unique solution and so directly applying ICA will not work in this case. In underdetermined situations, we typically need a stronger assumption than independence.

One successful approach for underdetermined BSS involves utilizing the sparse nature of speech [2–6], which relies on the fact that the time-frequency components of speech are near zero across most of the time-frequency bins. Since the time-frequency components of speech rarely overlap each other even when multiple speakers are speaking simultaneously, the main focus of this approach is how to design a time-frequency mask with which we can extract only the components of target speech from the mixture.

The signals observed at each microphone can be modeled as a convolutive mixture of source signals. To exploit the sparse nature of speech, we must convert it to a time-frequency representation. If we assume the use of a short-time Fourier transform (STFT) to obtain a time-frequency representation with a frame length sufficiently longer than the length of the impulse response from a source to a microphone, an observed signal can be approximated fairly well by an instantaneous mixture in the frequency domain. BSS based on this observation model is called frequency domain BSS. While frequency domain BSS allows for a fast implementation compared with BSS that uses a time domain convolutive mixture model, it requires us to solve an additional problem called the permutation alignment problem. That is, we must group together the separated components of different frequency bins that are considered to originate from the same source in order to construct a separated signal.

Motivated by the above, this paper proposes a novel BSS approach that simultaneously performs (1) an estimation of the number of sources, (2) source separation based on the sparseness of speech, and (3) permutation alignment.

2. MIXING MODEL

We first consider a situation where K source signals are captured by M microphones. Here, let $y_m(\omega,t) \in \mathbb{C}$ be the short-time Fourier transform (STFT) component observed at the m-th microphone, and $s_k(\omega,t) \in \mathbb{C}$ be the STFT component of the k-th source. $1 \le \omega \le \Omega$ and $1 \le t \le T$ are the frequency and time indices, respectively. If we assume that the length of the impulse response from a source to a microphone is sufficiently shorter than the frame length of the STFT, the observed signal can be approximated fairly well by an instantaneous mixture in the frequency domain:

$$\boldsymbol{y}(\omega,t)=\sum_{k=1}^{K}\boldsymbol{a}_k(\omega)s_k(\omega,t)+\boldsymbol{n}(\omega,t), \tag{1}$$

where $\boldsymbol{y}(\omega,t) = (y_1(\omega,t),\ldots,y_M(\omega,t))^{\mathsf{T}}$. $\boldsymbol{a}_k(\omega) = (a_{k,1}(\omega),\ldots,a_{k,M}(\omega))^{\mathsf{T}}$ is the frequency array response for source k, which is assumed to be time-invariant. $\boldsymbol{n}(\omega,t)$ is assumed to comprise all kinds of components that cannot be expressed by the instantaneous mixture representation (e.g., background noise and reverberant components).

We now exploit the sparseness of speech and assume that only one source is active in each time-frequency bin. By using $z_{\omega,t}\in\{1,\ldots,K\}$ to denote the (unknown) active source index at time-frequency bin (ω,t), (1) can be rewritten as

$$\boldsymbol{y}(\omega,t)=\boldsymbol{a}_{z_{\omega,t}}(\omega)s(\omega,t)+\boldsymbol{n}(\omega,t). \tag{2}$$

Notice that the superscript k is dropped from $s_k(\omega,t)$ in (2) as it is no longer necessary since we are assuming $s_k(\omega,t)=0$ for $k\neq z_{\omega,t}$. Namely, $s(\omega,t)$ signifies the component of an active source at time-frequency bin (ω,t). For convenience of notation, we hereafter use subscripts to indicate ω and t.

3. GENERATIVE MODEL

3.1. Generative process of observed signals

Here we describe the generative process of an observed signal on the basis of (2). Let us assume that the noise component $\boldsymbol{n}_{\omega,t}$ follows a complex normal distribution with mean $\mathbf{0}$ and covariance $\Sigma_\omega^{(n)}$. Then, from (2), $\boldsymbol{y}_{\omega,t}$ is also normally distributed such that

$$\boldsymbol{y}_{\omega,t}|\boldsymbol{a}_{1:K,\omega},s_{\omega,t},z_{\omega,t}\sim\mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{\omega,t};\boldsymbol{a}_{z_{\omega,t},\omega}s_{\omega,t},\Sigma_\omega^{(n)}), \tag{3}$$

conditioned on $\boldsymbol{a}_{1:K,\omega}=\{\boldsymbol{a}_{1,\omega},\ldots,\boldsymbol{a}_{K,\omega}\}$, $s_{\omega,t}$ and $z_{\omega,t}$, where $\mathcal{N}_{\mathbb{C}}(\boldsymbol{x};\boldsymbol{\mu},\Sigma)\propto\exp(-(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{H}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$.

3.2. Mixture of infinite sparse sources

We do not usually obtain any information about which source is active at each time-frequency bin. Thus, we regard $z_{\omega,t}$ as a latent variable and consider describing its generative process. If the number of sources is K, it would be natural to assume that the probability of choosing an index $z_{\omega,t}$ from the set of source indices, $\{1,\ldots,K\}$, can be described as a discrete distribution

$$z_{\omega,t}|\boldsymbol{\pi}\sim\mathrm{Discrete}(z_{\omega,t};\boldsymbol{\pi}), \tag{4}$$

where $\mathrm{Discrete}(x;\boldsymbol{y})=y_x$ (with $\boldsymbol{y}=(y_1,\ldots,y_K)$ and $\sum_{k=1}^K y_k=1$). The k-th element of $\boldsymbol{\pi}$ defines how likely the source index k is to be chosen. Since we do not also have any information about $\boldsymbol{\pi}$, we consider describing its generative process using a "symmetric" distribution. For the convenience of the following analysis, we assume that $\boldsymbol{\pi}$ has been generated from a symmetric Dirichlet distribution

$$\boldsymbol{\pi}\sim\mathrm{Dirichlet}(\boldsymbol{\pi};\alpha_0/K,\ldots,\alpha_0/K), \tag{5}$$

where $\mathrm{Dirichlet}(\boldsymbol{y};z_1,\ldots,z_K)\propto\prod_{k=1}^K y_k^{z_k-1}$. The shape of the Dirichlet distribution is governed by a concentration hyperparameter α_0.

Thus far, we have considered the case of a finite number K of sources. It can be shown that the limit of the above generative processes (3), (4) and (5) as K goes to infinity can be described explicitly as

$$\boldsymbol{\pi}\sim\mathrm{GEM}(\boldsymbol{\pi};\alpha_0), \tag{6}$$

$$z_{\omega,t}|\boldsymbol{\pi}\sim\mathrm{Discrete}(z_{\omega,t};\boldsymbol{\pi}), \tag{7}$$

$$\boldsymbol{y}_{\omega,t}|\boldsymbol{a}_{1:\infty,\omega}s_{\omega,t},z_{\omega,t}\sim\mathcal{N}_{\mathbb{C}}(\boldsymbol{y}_{\omega,t};\boldsymbol{a}_{z_{\omega,t},\omega}s_{\omega,t},\Sigma_\omega^{(n)}). \tag{8}$$

$\mathrm{GEM}(\boldsymbol{\pi};\alpha_0)$ is called the stick-breaking process [7], which is known as a constructive definition of the Dirichlet process [8] and is given by

$$v_k\sim\mathrm{Beta}(v_k;1,\alpha_0), \tag{9}$$

$$\pi_k=v_k\prod_{l=1}^{k-1}(1-v_l), \tag{10}$$

where $\mathrm{Beta}(y;z_1,z_2)\propto y^{z_1-1}(1-y)^{z_2-1}$. $\boldsymbol{\pi}\sim\mathrm{GEM}(\boldsymbol{\pi};\alpha_0)$ produces exponentially decaying weights in expectation. This means that the source with a larger index will be less likely to be active and thus simple models with fewer sources are favored, given observations.

3.3. Mixture of Direction-of-Arrivals (DOAs)

We describe here the generative process of the frequency response $\boldsymbol{a}_{k,\omega}$ of the mixing system.

So far we have treated $\boldsymbol{a}_{k,\omega}$ as an independent parameter across ω. If the index k indicates an identical source across ω, $\boldsymbol{a}_{k,\omega}$ will have a certain structure that can be described using the property of acoustic wave propagation. We thus expect that incorporating an appropriate constraint into $\boldsymbol{a}_{k,\omega}$ would help solve the permutation alignment problem through parameter inference. If each source is assumed to be located far from the microphones so that the signal can be treated approximately as a plane wave, the interchannel time difference between the microphones depends only on the direction of arrival (DOA) of the source. Since the time delay between two microphones corresponds to the phase difference of the frequency response of the microphone array, the complex array response can be expressed explicitly by using the DOAs of the source. Specifically, with $M=2$ microphones, the complex array response for a source at direction θ such that $0\leq\theta<2\pi$ is defined as a function of ω depending on θ

$$\boldsymbol{\ell}_{\theta,\omega}=\begin{bmatrix}1\\ e^{\jmath\omega D\cos\theta/C}\end{bmatrix}, \tag{11}$$

where $\jmath$ is the imaginary unit, D [m] is the distance between the two microphones, and C [m/s] is the speed of sound. If the DOA θ_k of source k is known, the frequency response $\boldsymbol{a}_{k,\omega}$ should be equal to $\boldsymbol{\ell}_{\theta_k,\omega}$. However, due to such factors as the plane wave assumption and the narrowband instantaneous mixture approximation, the actual frequency response $\boldsymbol{a}_{k,\omega}$ may diverge from the "ideal frequency response" $\boldsymbol{\ell}_{\theta_k,\omega}$

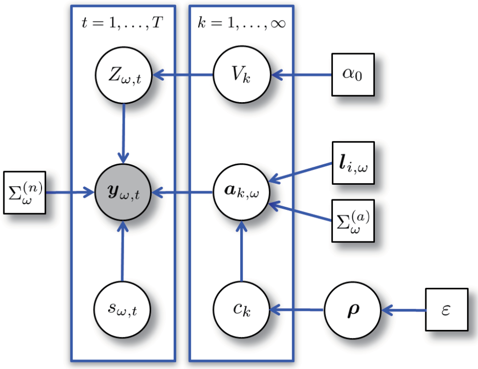

Fig. 1. Plate notation of the present generative model.

to some extent. One way to simplify the process of this kind of deviation is to assume a probability distribution on $\boldsymbol{a}_{k,\omega}$ with the expected value of $\boldsymbol{\ell}_{\theta_k,\omega}$. Here, we assume for convenience that $\boldsymbol{a}_{k,\omega}$ is generated from a complex normal distribution with mean $\boldsymbol{\ell}_{\theta_k,\omega}$. Note that we do not usually obtain any information about which direction each source is emanating from. Thus, we regard the DOA of each source as a latent variable and further consider describing its generative process. As explained in detail below, the entire generative process of $\boldsymbol{a}_{k,\omega}$ can then be described as a "mixture of DOAs".

Let us now introduce a discrete set of N possible directions, $\vartheta_1, \dots, \vartheta_N$, which are all assumed to be constants. For instance, consider defining ϑ_n as $\vartheta_n = (n-1)\pi/N$, $(n = 1, \dots, N)$ (dividing π in N equal angles). We assume that each source is emanating from one of these directions. First, we consider the generative process of the DOA θ_k of source k. For each source k, an index c_k of direction is drawn according to a discrete distribution $\boldsymbol{\rho} = (\rho_1, \dots, \rho_N)$

$$c_k | \boldsymbol{\rho} \sim \mathrm{Discrete}(c_k; \boldsymbol{\rho}). \tag{12}$$

By using c_k, θ_k is then given as

$$\theta_k = \vartheta_{c_k}. \tag{13}$$

As with 3.2, we assume that $\boldsymbol{\rho}$ has been generated from a symmetric Dirichlet distribution

$$\boldsymbol{\rho} \sim \mathrm{Dirichlet}(\boldsymbol{\rho}; \beta_0/M, \dots, \beta_0/M). \tag{14}$$

As mentioned above, the frequency response $\boldsymbol{a}_{k,\omega}$ is assumed to be generated from a complex normal distribution, given c_k,

$$\boldsymbol{a}_{k,\omega} | c_k \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{\ell}_{\theta_k,\omega}, \Sigma_\omega^{(a)}), \tag{15}$$

where $\Sigma_\omega^{(a)}$ denotes the covariance of the complex normal distribution. The entire generative model is described in Fig. 1.

4. APPROXIMATE POSTERIOR INFERENCE

In this section, we describe an approximate posterior inference algorithm for our generative model based on variational inference. The random variables of interest in our model are

$A = \{\boldsymbol{a}_{k,\omega}\}_{k,\omega}$: complex array response for source k,
$S = \{s_{\omega,t}\}_{\omega,t}$: component of active source at (ω, t),
$Z = \{z_{\omega,t}\}_{\omega,t}$: index of active source at (ω, t),
$V = \{v_k\}_k$: stick breaking proportion,
$C = \{c_k\}_k$: index of direction for source k,
$\boldsymbol{\rho} = (\rho_1, \dots, \rho_N)$: mixture weight for each DOA,

which we denote by Θ. Our goal is to compute the posterior $p(\Theta|Y)$ where $Y = \{\boldsymbol{y}_{\omega,t}\}_{\omega,t}$ is the set consisting of observed multichannel signals. By using the conditional distributions defined in 3.2 and 3.3, we can write the joint distribution $p(Y, \Theta)$ as

$$\begin{aligned} &p(Y, A, S, Z, V, C, \boldsymbol{\rho}) \\ &\quad = p(Y|A, S, Z)p(Z|V)p(V)p(A|C)p(C|\boldsymbol{\rho})p(\boldsymbol{\rho}), \end{aligned} \tag{16}$$

but to obtain the exact posterior $p(\Theta|Y)$, we must compute $p(Y)$, which involves many intractable integrals.

We can express this posterior variationally as the solution to an optimization problem:

$$\underset{q \in \mathcal{Q}}{\operatorname{argmin}} \, \mathrm{KL}(q(\Theta) \| p(\Theta|Y)), \tag{17}$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler (KL) divergence between its two arguments. Indeed, if we let $\mathcal{Q}$ be the family of all distributions over Θ, the solution to the optimization problem is the exact posterior $p(\Theta|Y)$, since KL divergence is minimized when its two arguments are exactly equal. Of course, solving this optimization problem is just as intractable as directly computing the posterior. Although it may appear that no progress has been made, having a variational formulation allows us to consider tractable choices of $\mathcal{Q}$ in order to obtain principled approximate solutions.

For our model, we define the set of approximate distributions $\mathcal{Q}$ to be those that factor as follows:

$$\mathcal{Q} = \big\{ q : q(A)q(S)q(Z)q(V)q(C)q(\boldsymbol{\rho}) \big\}. \tag{18}$$

To define $q(A)$, $q(V)$ and $q(C)$, we need to construct distributions on the infinite sets $\{v_1, v_2, \dots\}$, $\{\boldsymbol{a}_{1,\omega}, \boldsymbol{a}_{2,\omega}, \dots\}$ and $\{c_1, c_2, \dots\}$. For this approach to be tractable, we truncate the variational distribution at some value K^* by setting $q(v_{K^*} = 1) = 1$. The mixture proportions π_k for $k > K^*$ will thus be zero, and we can ignore $\boldsymbol{a}_{k,\omega}$ and c_k for $k > K^*$. In practice, we set K^* at a sufficiently large integer. It is important to emphasize that truncating the variational distribution does not mean that the true posterior itself is truncated. As the truncation level K^* becomes larger, the approximations to the true posterior become more accurate.

We now present an algorithm for solving the optimization problem described in (17) and (18). Unfortunately, the optimization problem is non-convex, and it is intractable to find the global optimum. However, we can use a simple coordinate ascent algorithm to find a local optimum. The algorithm optimizes one factor in the mean-field approximation of the posterior at a time while fixing all other factors. The mean-field update equations for the variational distributions are given in the following form:

$$\hat{q}(A) = \prod_{k,\omega} \mathcal{N}_{\mathbb{C}}(\boldsymbol{a}_{k,\omega}; \boldsymbol{m}_{k,\omega}, \Gamma_{k,\omega}), \tag{19}$$

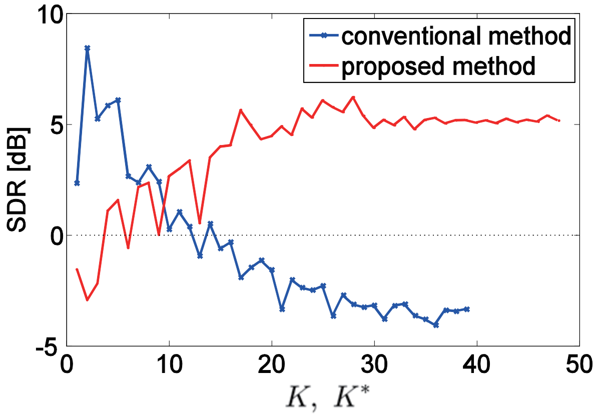

Fig. 2. SDRs obtained with conventional and present methods with different K and K^* settings.

$$\hat{q}(S) = \prod_{\omega,t} \mathcal{N}_{\mathbb{C}}(s_{\omega,t}; \mu_{\omega,t}, \sigma^2_{\omega,t}), \tag{20}$$

$$\hat{q}(Z) = \prod_{\omega,t} \mathrm{Discrete}(z_{\omega,t}; \boldsymbol{\phi}_{\omega,t}), \tag{21}$$

$$\hat{q}(V) = \prod_{k} \mathrm{Beta}(v_k; \gamma_{k,0}, \gamma_{k,1}), \tag{22}$$

$$\hat{q}(C) = \prod_{k} \mathrm{Discrete}(c_k; \boldsymbol{\psi}_k), \tag{23}$$

$$\hat{q}(\boldsymbol{\rho}) = \prod_{k} \mathrm{Dirichlet}(\boldsymbol{\rho}; \zeta_{k,1}, \ldots, \zeta_{k,M}). \tag{24}$$

5. EXPERIMENT

We evaluated the performance of the proposed method in terms of source separation ability.

We used a stereo speech signal with a sampling rate of 16kHz as a test signal, which we obtained by mixing three speech signals [9] (two female and one male speakers) using a measured room impulse response [10] (in which the distance between the microphones was 5 cm and the reverberation time was 0 ms). The three sources were spaced 30 degrees apart. To compute the STFT components of the observed signal, the STFT frame length was set at 64 ms and a Hamming window was used with an overlap length of 16ms. $\Sigma^{(n)}_{\omega}$ and $\Sigma^{(a)}_{\omega}$ were set respectively at $\boldsymbol{I}$ and $10^{-1.5} \times \boldsymbol{I}$. M was set at 180. All the variational parameters were initialized randomly. The variational inference algorithm was run for 100 iterations. After convergence, each separated signal was obtained by multiplying $\mu_{\omega,t}$ by $\phi_{k,\omega,t}$. We chose Sawada's method described in [11] as a comparison. In this method, the number of sources must be specified manually. The following results report the performance in terms of the Signal-to-Distortion Ratio (SDR) [12]. The SDR is expressed in decibels (dB), and a higher SDR indicates superior quality.

The present method was tested with various settings of the truncation level K^*. As for Sawada's method, it was tested with various settings of the assumed number K of sources. Fig. 2 shows the average SDRs obtained with Sawada's and the present methods with various K and K^* settings. As expected, the performance of the present method improves with increasing K^*, while that of Sawada's method deteriorates significantly when the assumed number of sources departs from the actual number.

6. CONCLUSION

This paper aimed at developing a BSS algorithm that works well even when the number of sources is unknown and proposed a novel BSS approach that simultaneously performs an estimation of the number of sources, source separation based on the sparseness of speech, and permutation alignment. We confirmed experimentally that reasonably good separations were obtained with the present method without specifying the number of sources.

7. REFERENCES

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

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

[3] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Adv. Neural Info. Process. Syst.*, 2006, pp. 953–960.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[5] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, and T. Morita, "Real-time implementation of two-stage blind source separation combining SIMO-ICA and binary masking," in *Proc. IWAENC '05*, 2005, pp. 229–232.

[6] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. WASPAA '07*, 2007, pp. 147–150.

[7] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[8] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[9] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Trans. Speech Communication*, pp. 357–363, 1990.

[10] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. LREC '00*, 2000, pp. 965–968.

[11] 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, 2010.

[12] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, pp. 1462–1469, 2006.