BLIND SOURCE SEPARATION OF MIXED SPEECH IN A HIGH REVERBERATION ENVIRONMENT

Keiju Iso[†‡] *Shoko Araki*[‡] *Shoji Makino*[†]
Tomohiro Nakatani[‡] *Hiroshi Sawada*[‡] *Takeshi Yamada*[†] *Atsushi Nakamura*[‡]

[†]University of Tsukuba
1-1-1 Tennoudai, Tsukuba, Ibaraki, 305-8573, Japan
Tel: +81-29-853-6564, email: iso@mmlab.cs.tsukuba.ac.jp
[‡]NTT Communication Science Laboratories, NTT Corporation,
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237, Japan

ABSTRACT

Blind source separation (BSS) is a technique for estimating and separating individual source signals from a mixed signal using only information observed by each sensor. BSS is still being developed for mixed signals that are affected by reverberation. In this paper, we propose combining the BSS method that considers reverberation proposed by Duong *et al.* with the BSS method reported by Sawada *et al.*, which does not consider reverberation, for the initial setting of the EM algorithm. This proposed method assumes the underdetermined case. In the experiment, we compare the proposed method with the conventional method reported by Duong *et al.* and that reported by Sawada *et al.*, and demonstrate the effectiveness of the proposed method.

Index Terms— blind source separation, reverberation, model parameter estimation

1. INTRODUCTION

Blind source separation (BSS) is a technique for separating a mixed signal using only information observed by individual sensors, and estimating each source signal. BSS techniques for speech signals are expected to find many applications, such as hands free TV conferencing systems.

Many methods such as independent component analysis [1], and sparse component analysis [2] [3], have already been developed and have performed well for mixed signals under low reverberation conditions. However, BSS for mixed signals with strong reverberation needs to be developed because most of previous methods assume an anechoic environment. In this paper, first of all, we discuss a source separation method that uses model parameter estimation considering the reverberation described by Duong *et al.* [4], which we call "the full-rank method". After that, we propose a method that solves a problem in the full-rank method.

In [4], Duong *et al.* assume that a source image including the spatial characteristics of reverberant source signals can be modeled by a complex Gaussian random variable, and propose considering full-rank spatial covariance matrices that model diffused late reverberation. They suggest a method of BSS by estimating the parameters of the covariance of the complex Gaussian random variable, and separating the individual sources using a Wiener filter designed using these parameters. To estimate the model parameters, they use the expectation maximization (EM) algorithm. However, its convergence depends heavily on its initial values. To obtain high separation performance, we have to find a good set of initial parameters. In this paper, for the initial value setting, we propose employing the results of the source separation method using frequency bin-wise clustering reported by Sawada *et al.*, which does not consider reverberation [5].

2. PROBLEM SETTING

2.1. Mixture

In a real environment, source signals observed using sensors can be modeled by convolutive mixtures as follows.

$$x_i(t) = \sum_{j=1}^{J} \sum_{\tau} h_{ij}(t) s_j(t-\tau) \quad (i = 1, \ldots, I) \qquad (1)$$

x_i is an signal observed by sensor i, h_{ij} is the impulse response from source j to sensor i. $s_j(j = 1, \ldots, J)$ are source signals. I and J are the numbers of sensors and source signals, respectively. With a vector notation, (1) can be expressed as follows:

$$\mathbf{x}(t) = \sum_{j=1}^{J} \sum_{\tau} \mathbf{h}_j(\tau) s_j(t-\tau) = \sum_{j=1}^{J} \mathbf{c}_j(t) \qquad (2)$$

where
$\mathbf{x}(t)=[x_1(t),x_2(t),\ldots,x_I(t)]^T$, $\mathbf{h}_j=[h_{1j},h_{2j},\ldots,h_{Ij}]^T$,
$\mathbf{c}_j=[c_1,c_2,\ldots,c_J]^T$.

$\mathbf{c}_j$ is a source image which is a sound signal containing spatial characteristics such as reverberation.

The goal of BSS is to obtain a separated signal $\mathbf{c}_j$ using only the observed signals $\mathbf{x}$. In this paper, we discuss the underdetermined problem ($I < J$), where I and J are assumed to be known.

This study employs a time-frequency domain representation of the signals. This is because a voice signal in the time-frequency domain is sparser than in the time domain [6], and convolution in the time domain can be approximated into a product in each frequency. An observed signal $\mathbf{x}$ in the time-frequency domain is modeled as follows.

$$\mathbf{x}(n,f)=\sum_{j=1}^{J}\mathbf{h}_j(f)s_j(n,f)\quad(j=1,\ldots,J) \tag{3}$$
$$=\sum_{j=1}^{J}\mathbf{c}_j(n,f)$$
$$\mathbf{c}_j(n,f)=\mathbf{h}_j(f)s_j(n,f) \tag{4}$$

Where $\mathbf{h}_j(f)$ represents the transfer function from source j to each sensor, $s_j(n,f)$ and $\mathbf{x}(n,f)$ and $\mathbf{c}_j(n,f)$ represent the source signal, the observed signal and the source image with a short-time Fourier transform signal respectively. n is the time frame index and f is the frequency.

2.2. Separation process

In this section, first, we outline the full-rank method [4], which estimates the source image using the model parameter estimation. This approach assumes that the source image $\mathbf{c}_j(n,f)$, which is a product of the transfer function $\mathbf{h}_j(f)$ and the source signal $s_j(n,f)$, is a random variable that has a complex Gaussian distribution with a covariance matrix $\mathbf{R}_{\mathbf{c}_j}$. The assumption is expressed in the following formula.

$$\mathrm{p}(\mathbf{c}_j;\mathbf{R}_{\mathbf{c}_j})=N(0,\mathbf{R}_{\mathbf{c}_j}) \tag{5}$$

From (4) the covariance matrix $\mathbf{R}_{\mathbf{c}_j}$ can be expressed as follows.

$$\mathbf{R}_{\mathbf{c}_j}(n,f)=\mathbf{c}_j(n,f)\mathbf{c}_j^H(n,f)$$
$$=|s_j(n,f)|^2\mathbf{h}_j(f)\mathbf{h}_j^H(f)$$
$$=v_j(n,f)\mathbf{R}_j(f) \tag{6}$$

From (6), the covariance matrix $\mathbf{R}_{\mathbf{c}_j}$ can be divided into time-dependent signal variance v_j and time-invariant covariance matrix $\mathbf{R}_j$ representing the spatial characteristics of sounds. Here if the observed signal consists only of the direct sound, the rank of a covariance matrix $\mathbf{R}_j$ will be one, but considering reverberation, the full-rank method assumes that $\mathbf{R}_j$ is full rank [4].

We estimate the model parameters $\theta=\{v_j,\mathbf{R}_j\}$ by maximizing the likelihood function as follows.

$$\mathrm{P}(\{\mathbf{x}\},\{\mathbf{c}_j\},\theta)=\prod_n\prod_f\mathrm{P}(\mathbf{x}\mid\{\mathbf{c}_j\},\theta)\prod_j\mathrm{P}(\mathbf{c}_j\mid\theta) \tag{7}$$

With this likelihood function, we iteratively update the model parameters v_j, $\mathbf{R}_j$ and the source image $\mathbf{c}_j$ using the EM algorithm.

Each step of the EM algorithm is calculated as follows.

Initialization: In [4], the initial values of the model parameters are as follows.

$$\mathbf{R}_j^{\mathrm{init}}(f)=\frac{1}{|C_j|}\sum_{\mathbf{x}(n,f)\in C_j}\mathbf{x}(n,f)\mathbf{x}(n,f)^H \tag{8}$$
$$v_j^{\mathrm{init}}(n,f)=1 \tag{9}$$

where $|C_j|$ denotes the total number of samples in cluster C_j, where each cluster member corresponds to the time-frequency component of each separated source $\hat{\mathbf{c}}_j$. The full-rank method calculates C_j by clustering $\mathbf{x}(n,f)$ with the hierarchical clustering approach [4].

E-step: Update Wiener filter $\mathbf{W}_j$ for separation and source image estimation $\hat{\mathbf{c}}_j$.

$$\mathbf{W}_j(n,f)=\mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f) \tag{10}$$
$$\hat{\mathbf{c}}_j(n,f)=\mathbf{W}_j(n,f)\mathbf{x}(n,f) \tag{11}$$
$$\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)=\hat{\mathbf{c}}_j(n,f)\hat{\mathbf{c}}_j^H(n,f)+(\mathbf{I}-\mathbf{W}_j(n,f))\mathbf{R}_{\mathbf{c}_j}(n,f) \tag{12}$$

where,

$$\mathbf{R}_{\mathbf{c}_j}(n,f)=v_j(n,f)\mathbf{R}_j(f) \tag{13}$$
$$\mathbf{R}_{\mathbf{x}}(n,f)=\sum_{j=1}^{J}v_j(n,f)\mathbf{R}_j(f) \tag{14}$$

M-step: Update model parameters v_j and $\mathbf{R}_j$.

$$v_j(n,f)=\frac{1}{I}\mathrm{trace}(\mathbf{R}_j^{-1}(f)\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f)) \tag{15}$$

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\hat{\mathbf{R}}_{\mathbf{c}_j}(n,f) \qquad (16)$$

With the iteration of these algorithms, separated signals are obtained by formula (11). Obtained signals are transformed back to the time domain by inverse short-time Fourier transformation.

3. PROPOSED METHOD

Instead of using hierarchical clustering to initialize the EM algorithm as in [4], our proposed method applies the source separation method which uses the bin-wise classification [5]. Figure 1 shows a block diagram of the proposed algorithm. In our proposed approach, at the initialization stage, we utilize the bin-wise classification method. In [5], the cluster C_j is estimated according to information on the vectors $\mathbf{h}_j$ estimated from the normalized norm of $\mathbf{x}(n,f)$ and determining the most dominant source for each time *n*. Since this method is known to have high separation performance under low reverberation conditions*, it is expected to provide us with good initial values of C_j. The proposed method utilizes this cluster C_j for the initial value calculation of $\mathbf{R}_j^{\text{init}}$ (8). The initial value v_j^{init} is set in the same way as (9). In Fig. 1, the model parameter estimation and Wiener filter estimation are executed in the same way as (10)-(16). The permutation of the frequency components of the separated signals were aligned with the method described in [7].

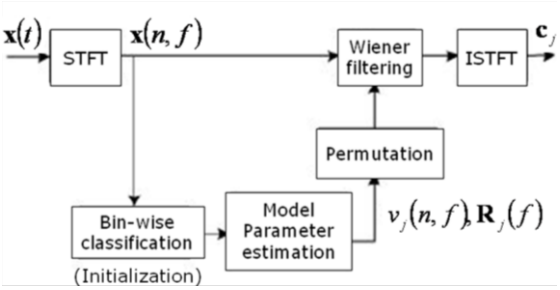

Fig. 1: The flow of algorithm

4. EXPERIMENT AND EVALUATION

4.1. Experiments: Comparison with two conventional methods

To evaluate the effectiveness of the proposed method, we compare its performance with that of the full-rank method with initialization (8) (9) [4], and with the frequency bin-wise clustering method [5].

In the experiments, we separated two-channel mixed audio signals with three speakers including reverberation, and measured the objective evaluation scores of the obtained separated signals. The objective evaluation scores include four measures proposed in [8].

- SDR (Signal to Distortion Ratio)
- ISR (source Image to Spatial distortion Ratio)
- SIR (Source to Interference Ratio)
- SAR (Sources to Artifacts Ratio)

They are expressed in dB, and higher numbers, indicate better performance.

4.2. Experimental condition

A. Experiment 1: Comparison with the full-rank method

Using audio data from SiSEC*2010, we compared the results obtained with the proposed method with those reported by Duong which were published in SiSEC2010. The reverberation time of the room was 250 ms, and the distance between two microphones was 5 cm. The sampling frequency was 16 kHz, the FFT size was 2048 and the overlap was 256. The proposed method was iterated 20 times.

B. Experiment 2: Comparison with the bin-wise clustering method

We separated eight seconds of audio signal mixed with three speakers, and 8 kHz sampling. We set the FFT size 1024. We evaluated the separation results for 1, 5, 10, 20, 30, 40, and 50 iterations of the proposed method, under six reverberation conditions (130 ~ 450 ms). We averaged the calculated values for six combinations of speakers.

4.3. Experimental results

The results of experiments 1 and 2 are shown in Tables 1, 2 and Figure 2 respectively.

Table 1: Results of experiment 1

Algorithm		SDR	ISR	SIR	SAR
Full-rank method	sim1	0.1	6.0	-0.5	7.8
	sim2	5.2	7.6	9.1	11.6
	sim3	3.0	4.8	5.7	8.0
Bin-wise clustering	sim1	7.0	10.9	13.2	8.9
	sim2	5.4	12.5	8.7	7.9
	sim3	8.5	12.7	15.0	10.8

* SiSEC: Signal Separation Evaluation Campaign (http://sisec.wiki.irisa.fr/tiki-index.php)

Proposed	sim1	7.8	11.1	11.8	11.6
	sim2	6.4	10.4	9.3	10.1
	sim3	9.2	12.9	13.2	13.7

Table 2: Results of experiment 2

Algorithm / Iteration	SDR	ISR	SIR	SAR
Bin-wise clustering	6.8	12.3	11.9	8.9
Proposed / 1	5.4	8.7	6.0	13.4
5	5.9	9.4	8.0	11.44
10	7.1	10.9	10.1	11.36
20	7.72	12.1	11.1	11.65
30	7.68	12.5	11.2	11.76
40	7.8	12.82	11.4	11.83
50	7.9	12.84	11.5	11.80

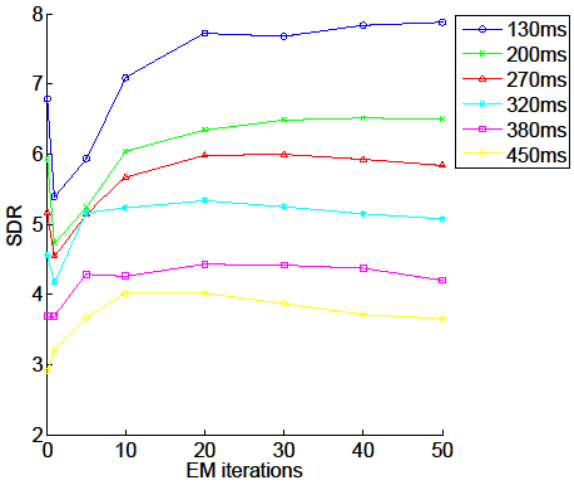

Fig. 2: SDR of the proposed method in Experiment 2 for six reverberation conditions

In Table 1, we also show the results of the bin-wise clustering method under the same conditions for reference. The reverberation time is 130 ms in Table 2. Figure 2 shows the SDR for each number of iterations of the proposed method in experiment 2, for six reverberation times. Iteration 0 indicates the results obtained using the initial value with the bin-wise clustering method. In Table 1, "sim" denotes "source image." Tables 1 and 2 (Fig. 2) show that the SDR representing overall distortion increases with the proposed method. This indicated the effectiveness of the proposed method.

4.4. Discussion

We discuss the way in which the performance was improved by the proposed method compared with the two conventional methods.

From Table 1, the SDR was poor with the full-rank method and we can obtain a significant improvement by setting the initial value in the proposed method. Table 2 shows that although The SIR with the proposed method is slightly lower than that of [5], the overall SDR performance is improved due to the SAR improvement. From this result, we can say that the sound quality is improved with the proposed method while maintaining separation performance, and that the overall performance is improved compared with the bin-wise clustering method. We also confirmed the quality of the separated sounds provided by each method by using an informal listening test and the results were consistent with the objective results.

Figure 2 shows that the SDR improvement for longer reverberation was small compared with the bin-wise clustering method which does not consider reverberation. Therefore, we may be able to improve the performance if we can model the reverberant signals more appropriately.

5. CONCLUSIONS

In this paper, we proposed a source separation method that combines a model parameter estimation method considering reverberation, and an initial value setting based on results obtained with a frequency bin-wise clustering approach. We confirmed the effectiveness of the proposed method experimentally.

6. REFERENCES

[1] S. Makino, S. Araki, R. Mukai, and H. Sawada, "Audio source separation based on independent component analysis," in *ISCAS 2004*, pp. V-668 – V-671, May 2004.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," in *Signal Processing, vol. 87*, pp. 1833-1847, 2007.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking", in *Signal Processing, vol. 52*, pp. 1830-1847, 2004.

[4] N. Q. K. Duong, E. Vincent and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model", *IEEE Trans. on ASLP*, vol. 18, no. 7, pp. 1830-1840, Sept. 2010.

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

[6] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time-Fourier transform, " in *ICA 2000*, pp.87-92, Jun. 2000.

[7] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in Frequency-domain BSS," in *ISCAS 2007*, pp. 3247-3250, May 2007.

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