NEW ANALYTICAL UPDATE RULE FOR TDOA INFERENCE FOR UNDERDETERMINED BSS IN NOISY ENVIRONMENTS

Takuro Maruyama*†*, Shoko Araki*†*, Tomohiro Nakatani*†*, Shigeki Miyabe***, Takeshi Yamada***,*
Shoji Makino* *and Atsushi Nakamura*†

*Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennoudai, Tsukuba-shi, Ibaraki 305-8573, Japan
†NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: maruyama@mmlab.cs.tsukuba.ac.jp

ABSTRACT

In this paper, we propose a new technique for sparseness-based underdetermined BSS that is based on the clustering of the frequency-dependent time difference of arrival (TDOA) information and that can cope with diffused noise environments. Such a method with an EM algorithm has already been proposed, however, it required a time-consuming exhaust search for TDOA inference. To remove the need for such an exhaust search, we propose a new technique by focusing on a stereo case. We derive an update rule for analytical TDOA estimation. This update rule eliminates the need for the exhaustive TDOA search, and therefore reduces the computational load. We show experimental results for separation performance and calculation time in comparison with those obtained with the conventional approach. Our reported results validate our proposed method, that is, our proposed method achieves high performance without a high computational cost.

Index Terms— Underdetermined blind source separation, EM algorithm, speech sparseness, time-frequency masking

1. INTRODUCTION

Blind Source Separation (BSS) has been intensively investigated since this problem setting matches a real environment very well. With overdetermined BSS, the source separation can be performed satisfactorily, especially in a clean environment, for example by using Independent Component Analysis (ICA). To be able to handle a more realistic situation, however, we must consider the underdetermined case, where there are fewer sensors than sources.

Many methods have been proposed for underdetermined BSS [1, 2, 3, 4, 5]. Most of there methods employ the sparseness characteristics of the source signals. The energy of the target signal concentrates on certain time-frequency points, which defines sparseness. In particular with speech in the time-frequency domain, it is approximately true owing to such factors as formants, harmonic structures, and non-stationarity. If the time-frequency components of each source signal are sparse, then these components rarely overlap even when many sources are mixed together [6], and it can be assumed that at most one source is active at each time-frequency slot. Time-frequency masking utilizes this sparseness assumption, and aims at extracting the time-frequency components dominated by each source. Such a time-frequency mask can be built by performing a clustering operation on the source location information, such as the TDOA, at all the time-frequency slots [1, 7]. However, most of the separation methods based on the clustering of the source location information tend to be weak with respect to reverberation and background noises.

Recently, Izumi et al. [8] proposed a BSS method to overcome this issue. They also relied on the TDOA for the clustering, however, their method achieved high performance even under noisy and reverberant conditions because they considered the noise and reverberation in their microphone observation model. The clustering was performed by using the EM algorithm, however, there is a serious problem with this approach, that is, it cannot analytically update the TDOA parameter. In other words, their method required a time-consuming exhaust search for the TDOA inference. This problem has precluded wide use of this approach.

To remove the need for such an exhaust search, we propose a new efficient BSS algorithm. We derive an update rule for TDOA parameter estimation in Izumi's algorithm in an analytical way. The new update rule eliminates the need for the exhaust search of TDOA, and is therefore expected to reduce the computational load. Our experimental results show that the proposed method can achieve the comparable performance by greatly reducing the computational time under noisy and reverberant conditions.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

2. CONVENTIONAL METHOD

This section outlines the method proposed by Izumi et al. [8] and its problem.

Let $\mathbf{x}_{f,t} = [x_{f,t,L}, x_{f,t,R}]^T$ be signals observed by two microphones represented in the time-frequency domain. If we assume that source signals are sufficiently sparse such that only one source signal is active at each time-frequency point, and each source signal is transferred as a plane wave, $\mathbf{x}_{f,t}$ can be written as

$$\begin{bmatrix} x_{f,t,L} \\ x_{f,t,R} \end{bmatrix} = \begin{bmatrix} 1 \\ e^{j2\pi f \delta_k} \end{bmatrix} S_{f,t,k} + \begin{bmatrix} N_{f,t,L} \\ N_{f,t,R} \end{bmatrix} \quad (1)$$

$$\mathbf{x}_{f,t} = \mathbf{b}_{f,k} S_{f,t,k} + \mathbf{N}_{f,t} \quad (2)$$

where k is the index of the source, $S_{f,t,k}$ is the spectrum of the source signal that is active at a time-frequency slot, $\mathbf{b}_{f,k}$ is the transfer function from the source to the microphones (δ_k is the TDOA between two microphones), and $\mathbf{N}_{f,t}$ is the observation error, which includes reverberation and background noise and is assumed to be independent of the source signals.

We assume that $\mathbf{N}$ is time-invariant and follows a Gaussian distribution with a zero mean and a covariance matrix $\sigma_f^2 \mathbf{V}_f$. Where σ^2 is the noise power, and $\mathbf{V}$ is given as follows for the diffused noise

$$\mathbf{V}(f) = \begin{bmatrix} 1 & \mathrm{sinc}(2\pi f D/c) \\ \mathrm{sinc}(2\pi f D/c) & 1 \end{bmatrix} \quad (3)$$

where c is the velocity of sound, and D represents the distance between the two microphones. The purpose of the conventional method [8] and this paper is to estimate the source signals $S_{f,t,k}$ solely from the mixed observation $\mathbf{x}_{f,t}$.

The likelihood function for the observation $\mathbf{x}_{f,t}$ is

$$p(\mathbf{x}_{f,t}|k,\theta) = \frac{1}{2\pi\sqrt{\sigma_f^2 |\mathbf{V}_f|}} \exp\left(-\frac{1}{2\sigma_f^2} \mathbf{N}_{f,t,k}^H \mathbf{V}_f^{-1} \mathbf{N}_{f,t,k} \right) \quad (4)$$

Let $\theta = \{\sigma_f^2, \delta_k, S_{f,t,k}\}$ be the parameter set. The log likelihood function is:

$$L = \sum_t \sum_f \log \sum_k p(\mathbf{x}_{f,t}|k,\theta) p(k|\theta) \quad (5)$$

where $p(k|\theta)$ is the mixing weight ($\sum_k p(k|\theta) = 1$).

In [8], this log likelihood is maximized with the EM algorithm. The parameters to be estimated are $\theta = \{\sigma_f^2, \delta_k, S_{f,t,k}\}$ where k is the hidden variable and the auxiliary function in this problem is

$$\begin{aligned} Q(\theta|\theta') &= \mathrm{E}\left[\log p(\mathbf{x}_{f,t};\theta)|\theta'\right] \\ &= \sum_t \sum_f \sum_k m_{k,f,t} \log p(\mathbf{x}_{f,t}|k,\theta) p(k|\theta) (6) \end{aligned}$$

where time-frequency mask $m_{k,f,t}$ is the posterior probability that source k is active at a t-f slot, and θ' is the parameter set obtained by the previous iteration. $m_{k,f,t}$ is updated by:

$$m_{f,t,k} = p(k|\mathbf{x}_{f,t,k}, \theta') = \frac{p(k|\theta') p(\mathbf{x}_{f,t}|k,\theta')}{\sum_k p(k|\theta') p(\mathbf{x}_{f,t}|k,\theta')} \quad (7)$$

The parameters σ_f^2 and $S_{f,t,k}$ are estimated by differentiating the auxiliary function with respect to each parameter, and setting them at zero,

$$\sigma_f^2 = \frac{1}{T} \sum_t \sum_k m_{f,t,k} \mathbf{N}_{f,t,k}^H \mathbf{V}_f^{-1} \mathbf{N}_{f,t,k} \quad (8)$$

$$S_{f,t,k} = \frac{\mathbf{b}_{f,k}^H \mathbf{V}_f^{-1} \mathbf{x}_{f,t}}{\mathbf{b}_{f,k}^H \mathbf{V}_f^{-1} \mathbf{b}_{f,k}}, \quad (9)$$

and the mixing weight $p(k|\theta)$ (where $\sum_k p(k|\theta) = 1$) is calculated by

$$p(k|\theta) = \frac{1}{TF} \sum_t \sum_f m_{f,t,k} \quad (10)$$

where T and F are the number of time frames and frequency bins, respectively.

In [8], as δ_k cannot be solved analytically, the update is performed by calculating $Q(\theta|\theta')$ for all the discretized δ_k and selecting δ_k that gives the maximum Q:

$$\delta_k = \mathrm{argmax}_{\delta_k} Q(\theta|\theta') \quad (11)$$

This update rule has two problems. One is that the inference of δ_k should always be discretized. Therefore, the optimum value tends not to be obtained when using these discretized values. The other problem is that this exhaust search requires a large computational cost. To overcome these problems, we derive an analytical update rule for estimating the TDOA parameter δ_k.

3. PROPOSED METHOD

In this section, we provide an analytical update rule for the TDOA parameter δ_k. By using the components of the vectors $\mathbf{x}$, $\mathbf{b}$ and the matrix $\mathbf{V}$, the likelihood function (4) can be rewritten as:

$$\begin{aligned} p(\mathbf{x}_{f,t}|k,\theta) &= \frac{1}{2\pi\sqrt{\sigma_f^2 |\mathbf{V}_f|}} \exp\big(C\big) \\ &\cdot \exp\Big(\frac{|\xi_{f,t,k}||S_{f,t,k}|}{\sigma_f^2 (1-\phi^2)} \cos(\psi_s - \psi_\xi - 2\pi f \delta_{f,k}) \Big) \quad (12) \end{aligned}$$

where $\phi_f = \mathrm{sinc}(2\pi f D/c)$, $\xi_{f,t,k} = [x_{f,t,R} - \phi_f(x_{f,t,L} - S_{f,t,k})]$, C is independent of $\delta_{f,k}$. ψ_{S_k} and ψ_{ξ_k} represent the phases of $S_{f,t,k}$ and $\xi_{f,t,k}$, respectively.

The last term,

$$\exp\Big(\frac{|\xi_{f,t,k}||S_{f,t,k}|}{\sigma_f^2(1-\phi^2)}\cos(\psi_s - \psi_\xi - 2\pi f\delta_{f,k})\Big), \qquad (13)$$

has the shape of the von Mises distribution [9]:

$$g(x|\kappa,\mu) \quad = \quad \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(x-\mu)} \qquad (14)$$

where $-\pi < x \leq \pi$, μ is the mean of the distribution ($-\pi < \mu \leq \pi$), $\kappa > 0$ is a concentration parameter, and $I_0(x)$ is a modified Bessel function of the first kind and order zero. That is, (13) means that the phase difference $\psi_s - \psi_\xi \approx \arg(x_{f,t,R}) - \arg(x_{f,t,L})$ has a von Mises distribution whose mean value corresponds to the frequency-dependent TDOA $\mu = 2\pi f\delta_{f,k}$, and the concentration parameter is the SNR related [1] value $\kappa = \frac{|\xi_{f,t,k}||S_{f,t,k}|}{\sigma_f^2(1-\phi^2)}$.

Therefore, we can derive the update rule for δ_k using a similar method with the mixture model of the von Mises distribution. However, because the cosine-part of (13) depends on the frequency f, we have to derive the update rule for $\delta_{f,k}$ at each frequency, which is different from the previous frequency independent update rule (11). By substituting (13) into (6), and by setting $\frac{\partial Q}{\partial \delta_{f,k}} = 0$, the update rule becomes:

$$2\pi f\delta_{f,k} = \arctan\frac{\sum_t m_{f,t,k}|\xi_{f,t,k}||S_{f,t,k}|\sin(\psi_{\xi_k} - \psi_{S_k})}{\sum_t m_{f,t,k}|\xi_{f,t,k}||S_{f,t,k}|\cos(\psi_{\xi_k} - \psi_{S_k})} \qquad (15)$$

It should be noted that the function $\arctan(x)$ is unique only if $-\pi/2 < x < \pi/2$. However, $2\pi f\delta_{f,k}$ can fall in the $-\pi$ to π range. Therefore, when $|x| \geq \pi/2$, we have to modify the estimated value by checking the inflection point of the auxiliary function. To accomplish this, we calculate the second order differential of the auxiliary function, and modify the values as follows:

- If $\delta_{f,k} < 0$, $\frac{\partial^2 Q}{\partial(\delta_{f,k}^2)} \geq 0$, then $2\pi f\delta_{f,k} \leftarrow 2\pi f\delta_{f,k} + \pi$
- If $\delta_{f,k} > 0$, $\frac{\partial^2 Q}{\partial(\delta_{f,k}^2)} \geq 0$, then $2\pi f\delta_{f,k} \leftarrow 2\pi f\delta_{f,k} - \pi$
- Otherwise, we do not modify $2\pi f\delta_{f,k}$

In summary, the proposed method estimates the parameters σ_f^2, $S_{f,t,k}$ and $m_{f,t,k}$ in the same ways as described in Section 2, and the TDOA parameter $\delta_{f,k}$ is calculated with the update rule (15).

4. EXPERIMENTS

4.1. Experimental conditions

We performed experiments with measured impulse responses in a room (Fig. 1) with a reverberation time of 130 ms. We used two microphones, whose spacing was 4 cm. The number of sources K was $K = 2$ ($70°$ and $150°$), or $K = 3$ ($30°, 70°$ and $150°$). Mixtures were formed by convolving the measured room impulse responses and 5-second English speech signals sampled at 8 kHz. The frame size and frame shift for STFT were 64 ms and 16 ms, respectively.

The noise we used in our experiments was the Gaussian noise of the zero mean and the covariance matrix of $\sigma_f^2\mathbf{V}_f$, where $\mathbf{V}_f$ was given by (3), and σ_f was determined so that the signal to noise ratio (SNR) with respect to source 1 has a preset value.

The performance was evaluated in terms of the signal to interference-plus-noise ratio (SINR) and the signal-to-distortion ratio (SDR) [2]. For each K, 10 speaker combinations were tested and the results were averaged.

We compared the performance and computational times of the conventional and proposed methods. For the parameter search with the conventional method, we changed the DOAs of the sources from $0°$ to $180°$ in increments of $1°$ and provided the corresponding TDOA δ_k. We then performed an exhaust search according to (11).

4.2. Results

Table 1 shows the experimental results, SINR, SDR and the required computational time (Intel Xeon X5650 2.67 GHz (6 Core), dual).

We can see from Table 1 that our proposed method achieves a comparable performance to the conventional approach and reduces the computational time by 1/10. This result shows that our proposed method can estimate the TDOA parameter δ_k without using an exhaust search as with the conventional method.

4.3. Influence of noise and reverberation

To investigate the influence of noise power and reverberation on separation performance, we also compared the performance and computational times of the conventional and the proposed methods by changing the SNR and reverberation time.

Table 2 shows the influence of the SNR on separation performance. The number of sources K was 3 ($30°, 70°$ and $150°$), and the other conditions were the same as those in Section 4.1. By adjusting the power of the noise σ_f^2, the SNR was set at 27.5 dB (= Table 1. (b)), 20 dB (Table 2. (a)), and 10 dB (Table 2. (b)).

Table 3 shows the influence of the reverberation time on separation performance. The number of sources K was 3

[1] When we assume that $\phi_f = 0$, $\frac{|\xi_{f,t,k}||S_{f,t,k}|}{\sigma_f^2(1-\phi^2)} = \frac{|S_{f,t,k}|^2}{\sigma_f^2}$, which is the SNR.

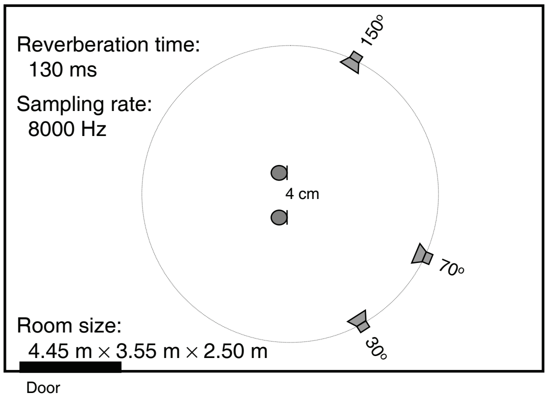

Fig. 1. Experimental setup

Table 1. Source separation results

(a) $K = 2$ $(70°$ and $150°)$

Method	SINR	SDR	Calculation Time
Conventional	15.3[dB]	6.7[dB]	184.1[s]
Proposed	14.8[dB]	7.7[dB]	16.3[s]

(b) $K = 3$ $(30°, 70°$ and $150°)$

Method	SINR	SDR	Calculation Time
Conventional	7.8[dB]	5.4[dB]	201.8[s]
Proposed	6.7[dB]	5.9[dB]	24.7[s]

Table 2. Performance under several SNR conditions

(a) $SNR = 20$ [dB]

Method	SINR	SDR	Calculation Time
Conventional	7.6[dB]	5.4[dB]	214.9[s]
Proposed	6.7[dB]	5.9[dB]	26.0[s]

(b) $SNR = 10$ [dB]

Method	SINR	SDR	Calculation Time
Conventional	5.3[dB]	5.4[dB]	231.0[s]
Proposed	6.1[dB]	5.9[dB]	29.6[s]

Table 3. Performance under several reverberation conditions

(a) reverberation time = 130 [ms]

Method	SINR	SDR	Calculation Time
Conventional	13.0[dB]	6.7[dB]	321.3[s]
Proposed	11.4[dB]	5.9[dB]	30.3[s]

(b) reverberation time = 300 [ms]

Method	SINR	SDR	Calculation Time
Conventional	7.0[dB]	2.9[dB]	422.9[s]
Proposed	6.3[dB]	2.8[dB]	42.1[s]

$(30°, 100°$ and $135°)$, and the other conditions were the same as those in Section 4.1. The reverberation time was set at 130 and 300 ms.

We can see from Tables 2 and 3 that our proposed method achieves comparable performance, reducing the computational time, regardless of the SNR and reverberation time.

5. CONCLUSION

This paper proposed a new method for sparse source separation in noisy and reverberant conditions. We provided an analytical update rule for estimating the TDOA parameter, which is estimated by using an exhaust search in the conventional approach. We confirmed that our proposed method can achieve comparable performance by reducing the computational time drastically, regardless of the power of noise. We plan to evaluate our method in a real environment, and to introduce a noise model appropriate for reverberation.

6. REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," IEEE Transactions on Signal Processing, Vol. 52, No. 7, pp 1830-1847, 2004.

[2] S. Araki and H. Sawada and R. Mukai and S. Makino, "Underdetermined Blind Sparse Source Separation for Arbitrarily Arranged Multiple Sensors," Signal Processing, Vol 77, No. 8, pp 1833-1847, 2007.

[3] M. Mandel and D. Ellis and T. Jebara, "An EM algorithm for Localizing Multiple Sound Sources in Reverberant Environments," *Proc. Neural Info. Proc. Sys.* , 2006.

[4] C. Fevotte and S. J. Godsill, "A Bayesian Approach for Blind Separation of Sparse Sources," IEEE Transactions on Speech and Audio Processing, Vol. 14, No. 6, pp 2174-2188, 2006.

[5] H. Sawada and S. Araki and S. Makino, "A Two-stage Frequency-domain Blind Source Separation Method for Underdetermined Convolutive Mixtures," *Proc. WASPAA2007.* , pp 139-142, 2007.

[6] S. Rickard and O. Yilmaz, "On the Approximate W-disjoint Orthogonality of Speech," *Proc. ICASSP*, Vol. I, pp. 529-532, 2002,

[7] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete BSS for Convolutive Mixtures Based on Hierarchical Clustering," *Proc. SAPA2004*, S13, 2004.

[8] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM Algorithm in Reverberant Environment," in *Proc. WASPAA2007*, pp. 147-150, 2007.

[9] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2008