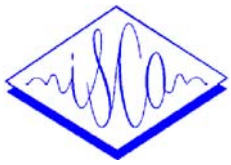

ITRW on Statistical and
Perceptual Audio Processing
ICC Jeju, Korea
October 3, 2004

Sound Source Localization and Separation Based on the EM Algorithm

Futoshi Asano and Hideki Asoh

Information Technology Res. Inst., AIST
Tsukuba, Japan
f.asano/h.asoh@aist.go.jp

Abstract

A method of sound localization using the EM algorithm has been proposed by Feder and Weinstein [1] and Miller and Fuhrmann [2]. In this paper, the signal separation aspect of this algorithm is analyzed and is extended so that it can be applied to separation of signals from moving sound sources.

1. Introduction

For using a speech interface in an everyday situation, separation of the target speech from environmental noise and other competing speech is indispensable. Various approaches including multi-microphone-based one such as adaptive beamforming and blind separation have been proposed (e.g., [3]). When sound sources do not move, these approaches perform well. When the sound sources move, however, the separation performance is usually degraded since the speed of adaptation is insufficient for dynamical change of source location.

[1] and [2] proposed a method of sound localization (*not separation*) based on the EM (Expectation-maximization) algorithm. In this method, a model of the covariance is introduced in the process of estimating the source location and the precision of this model is improved using expectation-maximization iteration. Due to this model-based approach, a higher performance of localization was achieved with a smaller amount of observation compared to other conventional localization method such as MUSIC [4]. This method is further developed by [5] for tracking of multiple moving targets by combining with Kalman-filter smoother.

An interesting feature of this method is that a signal separation mechanism is embedded in the sound localization process. In the present paper, this separation mechanism is focused on and is analyzed. Based on this analysis, the EM-based approach is then extended to the separation of sound from multiple moving sources.

2. Sound localization using the EM algorithm

In this section, the method of sound localization based on the EM algorithm is briefly reviewed to facilitate understanding of the following section.

2.1. Model of Signal

In this paper, a signal is treated in the frequency domain. The short-time Fourier transform (STFT) of the microphone input is denoted as $\mathbf{y}(\omega, t) = [Y_1(\omega, t)], \cdots, Y_M(\omega, t)]^T$ (input vector), where $Y_m(\omega, t)$ denotes the STFT of the mth microphone input at time t and frequency ω. Hereafter, the index of frequency ω is omitted for the sake of simplicity. The input vector for L directional signals plus background noise is modeled as

$$\mathbf{y}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{1}$$

Here, the matrix $\mathbf{A}$ consists of the location vector as $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_L]$. Each location vector $\mathbf{a}_l$ consists of a transfer function of the direct path from the lth source to the microphone as $\mathbf{a}_l = [A_{1,l}e^{-j\omega\tau_{1,l}}, \cdots, A_{M,l}e^{-j\omega\tau_{M,l}}]$ where $A_{m,l}$ and $\tau_{m,l}$ denote the gain and the time delay between the lth source and the mth microphone. The vector $\mathbf{s}(t)$ consists of the source spectrum as $\mathbf{s} = [S_1(t), \cdots, S_L(t)]^T$. The noise vector $\mathbf{n}(t)$ consists of background noise as $\mathbf{n} = [N_1(t), \cdots, N_M(t)]^T$. The noise is assumed to be 0-mean Gaussian noise. The symbol M denotes the number of microphones. It is assumed that $E[\mathbf{s}(t)\mathbf{s}^H(t)] \equiv \mathbf{K}_s = \mathrm{diag}(\gamma_1, \cdots, \gamma_L)$ and $E[\mathbf{n}(t)\mathbf{n}^H(t)] = \sigma\mathbf{I}$, where $\{\gamma_1, \cdots, \gamma_L\}$ denotes the power of signal $\mathbf{s}(t)$ while σ denotes the power of noise $\mathbf{n}(t)$.

2.2. EM Algorithm

In the EM-based approach, the input vector is decomposed into that corresponding to each sound source, $\mathbf{x}_l(t)$, as

$$\mathbf{y}(t) = \sum_{l=1}^{L} \mathbf{x}_l(t) = \mathbf{H}\mathbf{x}(t), \tag{2}$$

where $\mathbf{x}_l(t) = \mathbf{a}_l S_l(t) + \mathbf{n}_l(t)$, $\mathbf{x}(t) = [\mathbf{x}_1^T(t), \cdots, \mathbf{x}_L^T(t)]^T$, and $\mathbf{H} = [\mathbf{I}, \cdots, \mathbf{I}]$. The matrix $\mathbf{I}$ denotes the identity matrix. The symbol $\mathbf{n}_l(t)$is an arbitrary decomposition of the noise vector $\mathbf{n}(t)$ and satisfies $\sum_{l=1}^{L} \mathbf{n}_l(t) = \mathbf{n}(t)$ and $E[\mathbf{n}_l(t)\mathbf{n}_l^H(t)] = \frac{\sigma}{L}\mathbf{I}$. A set of decomposed input vectors, $\mathbf{X} = [\mathbf{x}(1), \cdots, \mathbf{x}(N)]$, is termed *complete data* in the EM Algorithm. Using this decomposition, the log-likelihood

function becomes

$$L_c(\Theta, \mathbf{K}_s; \mathbf{X}) = \sum_{l=1}^{L} \left[-N \log\det \mathbf{K}_{xl} - \sum_{n=1}^{N} \mathbf{x}_l^H(t) \mathbf{K}_{xl}^{-1} \mathbf{x}_l(t) \right], \tag{3}$$

where

$$\mathbf{K}_{xl} = \gamma_l \mathbf{a}_l \mathbf{a}_l^H + \frac{\sigma}{L}\mathbf{I} \tag{4}$$

The covariance of the observation $\mathbf{y}(t)$ is written using (4) as $\mathbf{K}_y \equiv E[\mathbf{y}(t)\mathbf{y}^H(t)] = \sum_{l=1}^{L} \mathbf{K}_{xl}$ The covariance matrices $\mathbf{K}_{xl}$ and $\mathbf{K}_y$ are termed *true covariance* hereafter. The parameter $\Theta = [\theta_1, \cdots, \theta_L]$ denotes the directions of sources.

The *sample covariance* of the complete data is defined as

$$\mathbf{C}_{xl} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{x}_l(t) \mathbf{x}_l^H(t) \tag{5}$$

In the E-step of the EM algorithm, the conditional expectation of $\mathbf{C}_{xl}$ is estimated. In the M-step, the log likelihood (3) is maximized using the conditional expectation of $\mathbf{C}_{xl}$. This procedure is summarized as follows.

E-Step:

$$\begin{aligned} \mathbf{C}_{xl}^p &\equiv E[\mathbf{C}_{xl}|\mathbf{C}_y; \hat{\mathbf{K}}_y^p] = \hat{\mathbf{K}}_{xl}^p - \hat{\mathbf{K}}_{xl}^p (\hat{\mathbf{K}}_y^p)^{-1} \hat{\mathbf{K}}_{xl}^p \\ &\quad + \hat{\mathbf{K}}_{xl}^p (\hat{\mathbf{K}}_y^p)^{-1} \mathbf{C}_y (\hat{\mathbf{K}}_y^p)^{-1} \hat{\mathbf{K}}_{xl}^p \qquad (6) \\ \hat{\mathbf{K}}_y^p &= \sum_{l=1}^{L} \hat{\mathbf{K}}_{xl}^p \qquad (7) \\ \hat{\mathbf{K}}_{xl}^p &= \hat{\gamma}_l^p \mathbf{a}(\hat{\theta}_l^p) \mathbf{a}(\hat{\theta}_l^p)^H + \frac{\sigma}{L}\mathbf{I} \qquad (8) \end{aligned}$$

M-Step:

$$\begin{aligned} \hat{\theta}_l^{p+1} &= \arg\max_{\theta_l} \frac{\mathbf{a}^H(\theta_l) \mathbf{C}_{xl}^p \mathbf{a}(\theta_l)}{|\mathbf{a}(\theta_l)|^2} \qquad (9) \\ \hat{\gamma}_l^{p+1} &= \frac{\mathbf{a}^H(\hat{\theta}_l^{p+1}) \mathbf{C}_{xl}^p \mathbf{a}(\hat{\theta}_l^{p+1})}{|\mathbf{a}(\hat{\theta}_l^{p+1})|^4} \qquad (10) \end{aligned}$$

The symbol $\hat{\cdot}$ indicates the estimate in the EM algorithm. The covariance matrices $\hat{\mathbf{K}}_{xl}$ and $\hat{\mathbf{K}}_y$ are termed *model covariance* hereafter to distinguish them from the true covariance. The superscript $\cdot^p$ denotes the iteration number of the EM algorithm. A brief derivation of the expected value of the sample covariance $E[\mathbf{C}_{xl}]$ is shown in Appendix so that it can be used for the analysis of signal separation in the EM algorithm.

3. Analysis of the EM-based Approach

3.1. Beamformer

In this subsection, some basic knowledge of the beamformer, which is used for the analysis, is reviewed [6]. The delay-and-sum (DS) beamformer, the maximum likelihood (ML) beamformer and the minimum variance (MV) beamformer that focuses on the Jth sound source is given as

$$\begin{aligned} z(t) &= \mathbf{w}_{DS}^H \mathbf{y}(t) = \frac{\mathbf{a}_J^H}{\mathbf{a}_J^H \mathbf{a}_J} \mathbf{y}(t) \qquad (11) \\ z(t) &= \mathbf{w}_{ML}^H \mathbf{y}(t) = \frac{\mathbf{a}_J^H \mathbf{K}_n^{-1}}{\mathbf{a}_J^H \mathbf{K}_n^{-1} \mathbf{a}_J} \mathbf{y}(t) \qquad (12) \\ z(t) &= \mathbf{w}_{MV}^H \mathbf{y}(t) = \frac{\mathbf{a}_J^H \mathbf{C}_y^{-1}}{\mathbf{a}_J^H \mathbf{C}_y^{-1} \mathbf{a}_J} \mathbf{y}(t) \qquad (13) \end{aligned}$$

The vectors, $\mathbf{w}_{DS}$, $\mathbf{w}_{ML}$ and $\mathbf{w}_{MV}$ are their coefficients. In the ML beamformer, covariance $\mathbf{K}_n$ is defined as the covariance without the target signal and is defined as

$$\mathbf{K}_n = \sum_{l \neq J} \gamma_l \mathbf{a}_l \mathbf{a}_l^H + \sigma \mathbf{I} \tag{14}$$

In the ML and MV beamformers, notches are made in the direction of sound sources other than Jth sound source based on the information included in the covariance $\mathbf{K}_n$ and $\mathbf{C}_y$. This effect is termed the adaptation and the separation performance is usually higher than that of the DS beamformer. In Eqs.(11), (12) and (13), the denominators of the coefficient vectors are the normalization factors. The effect of signal separation is included in their numerators.

3.2. Signal separation in the EM algorithm

In this sub-section, the effect of signal separation embedded in the EM algorithm is analyzed. From (35), the expected value of the sample covariance of $\mathbf{x}(t)$ is written as

$$E[\mathbf{C}_x] = \overbrace{(\mathbf{I} - \mathbf{G}\mathbf{H})\hat{\mathbf{K}}_x}^{Model} + \overbrace{\mathbf{G}\mathbf{C}_y\mathbf{G}^H}^{Observation} \tag{15}$$

In (15), the first term is related to the model covariance $\hat{\mathbf{K}}_x$ and the second term is related to the sample covariance $\mathbf{C}_y$ obtained from the observation $\{\mathbf{y}(1), \cdots, \mathbf{y}(N)\}$. From (30), the matrix $\mathbf{G}$ which appears in (15) has the following function:

$$\hat{\mathbf{x}}(t) = \mathbf{G}\mathbf{y}(t). \tag{16}$$

Since $\hat{\mathbf{x}}(t)$ is a separated observation, it can be known that $\mathbf{G}$ has a function of signal separation.

Using (37), (29) can be rewritten as

$$\mathbf{G} = \left[\hat{\mathbf{K}}_{x1} \hat{\mathbf{K}}_{x2} \cdots \hat{\mathbf{K}}_{xM} \right]^H \hat{\mathbf{K}}_y^{-1}. \tag{17}$$

Extracting the term related to the Jth source,

$$\mathbf{G}_J = \hat{\mathbf{K}}_{xJ} \hat{\mathbf{K}}_y^{-1} \tag{18}$$

The matrix $\mathbf{G}_J$ is termed the gain matrix hereafter. Substituting (4) into (18),

$$\mathbf{G}_J = \left[\hat{\gamma}_J \hat{\mathbf{a}}_J \hat{\mathbf{a}}_J^H + \frac{\sigma}{M}\mathbf{I} \right] \hat{\mathbf{K}}_y^{-1} \tag{19}$$

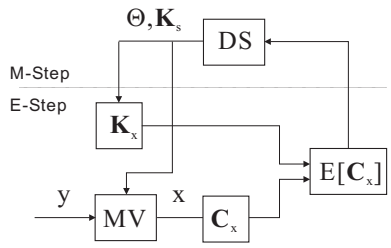

Figure 1: Block diagram of the EM-based sound localization.

For the sake of simplicity, assuming that the power of noise $\mathbf{n}(t)$ is small, i.e., $\hat{\gamma}_J >> \sigma/M$,

$$\mathbf{G}_J \simeq \hat{\gamma}_J \hat{\mathbf{a}}_J \hat{\mathbf{a}}_J^H \hat{\mathbf{K}}_y^{-1} = (\hat{\gamma}_J \hat{\mathbf{a}}_J)(\hat{\mathbf{a}}_J^H \hat{\mathbf{K}}_y^{-1}) \qquad (20)$$

By comparing (20) with (13), the latter term in (20), $\hat{\mathbf{a}}_J^H \hat{\mathbf{K}}_y^{-1}$, has the same form as that in the numerator in (13). From this, the term $\hat{\mathbf{a}}_J^H \hat{\mathbf{K}}_y^{-1}$ has a function of signal separation equivalent to the MV beamformer. The main difference between the MV beamformer (13) and (20) is that the sample covariance $\mathbf{C}_y$ is substituted by the model covariance $\hat{\mathbf{K}}_y$. The effect of this substitution is as follows:

- Mismatch of the parameter in the model covariance $\hat{\mathbf{K}}_y$ and the sample covariance $\mathbf{C}_y$.
- While cross-terms of the signals exist in $\mathbf{C}_y$, no cross-terms exist in $\hat{\mathbf{K}}_y$.

The mismatch of the parameters causes a mismatch of the direction of notches generated by the separation filter and the location of the noise source. This results in a deterioration in the separation performance. Regarding the cross-term, in the sample covariance $\mathbf{C}_y$, the cross-terms between different sources such as $\sum_{t=1}^{N}[\mathbf{a}_i s_i(t)][\mathbf{a}_j s_j(t)]^H (i \neq j)$ exist. When $N \to \infty$, these terms converge to 0 if the source signals are mutually uncorrelated. However, when N is small, the effect of the cross-terms in $\mathbf{C}_y$ is not negligible, resulting in a deterioration in the separation performance. Therefore, when the estimated parameters are precise, the separation performance of the gain matrix $\mathbf{G}_J$ is expected to be higher than that of the conventional MV beamformer.

On the other hand, the former term $\hat{\gamma}_J \hat{\mathbf{a}}_J$ has a function of recovering the microphone observation from the separated source signal and is not related to the signal separation.

3.3. Entire Function

As described above, the E-step of the EM algorithm has a function of signal separation. Next, the M-step is considered. The spatial spectrum estimated by the DS beamformer is given as [6]

$$P(\theta) = \mathbf{w}_{DS}^H(\theta) \mathbf{C}_y \mathbf{w}_{DS}(\theta) = \frac{\mathbf{a}^H(\theta) \mathbf{C}_y \mathbf{a}(\theta)}{|\mathbf{a}(\theta)|^4} \qquad (21)$$

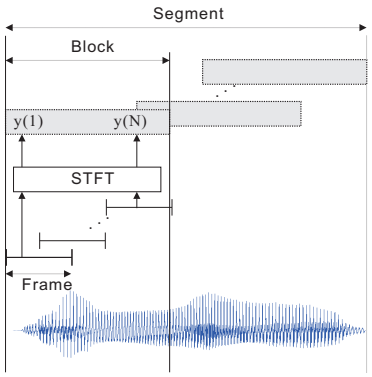

Figure 2: Relation of frame, block and segment.

By comparing this with (9) and (10), it can be seen that, in the M-step, the DS beamforming is performed on the separated observation $\mathbf{x}_l(t)$.

Figure 1 summarizes the entire function. In the E-step, using the estimated parameter $\hat{\Theta}$, the MV beamformer is constructed, and the separated observation $\mathbf{x}_l(t)$ is estimated. Using $\mathbf{x}_l(t)$, the sample covariance $\mathbf{C}_{xl}$ is then estimated. Finally, using the sample covariance $\mathbf{C}_{xl}$ and the model covariance $\hat{\mathbf{K}}_{xl}^p$, the expected value $E[\mathbf{C}_{xl}|\mathbf{C}_y; \hat{\mathbf{K}}_{xl}^p]$ is obtained. In the M-step, using the expected value $E[\mathbf{C}_{xl}|\mathbf{C}_y; \hat{\mathbf{K}}_{xl}^p]$, the DS beamformer is constructed, and the parameters of the signals, $\hat{\Theta}$ and $\hat{\mathbf{K}}_s$, are re-estimated.

4. Proposed Method for Sound Source Tracking and Separation

4.1. Sound Separation

As described in Section 3.2, the gain matrix $\mathbf{G}_J$ has a function of signal separation equivalent to that of the MV beamformer. In this section, this property is utilized for signal separation. Instead of directly using the gain matrix $\mathbf{G}_J$, a linear beamformer is explicitly constructed by utilizing the intermediate variables in the EM algorithm.

The ML and MV beamformers described in Section 3.1 can be generalized as follows:

$$\mathbf{w} = \frac{\mathbf{R}^{-1}\hat{\mathbf{a}}_J}{\hat{\mathbf{a}}_J^H \mathbf{R}^{-1} \hat{\mathbf{a}}_J} \qquad (22)$$

Here, $\hat{\mathbf{a}}_J$ is the location vector of the Jth source (target source) estimated in the sound localization. When $\mathbf{R} = \mathbf{C}_y$ or $\mathbf{R} = \mathbf{K}_n$ is employed, (22) becomes the original MV or ML beamformers, respectively. In this paper, these beamformers are approximated by using the intermittent variables

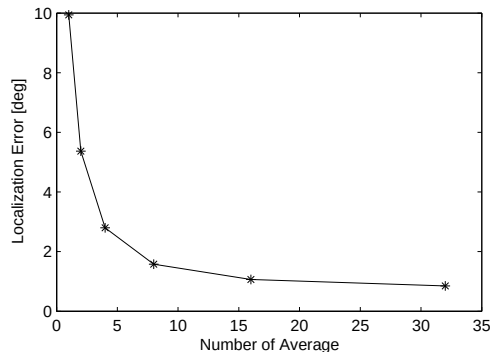

Figure 3: Estimation error as a function of N.

in the EM algorithm as follows:

$$\mathbf{R} = \hat{\mathbf{K}}_y^p \tag{23}$$

$$\mathbf{R} = E[\mathbf{C}_y] = \sum_{l=1}^{L} \mathbf{C}_{xl}^p \tag{24}$$

$$\mathbf{R} = \hat{\mathbf{K}}_n = \sum_{l \neq J} \hat{\mathbf{K}}_{xl}^p \tag{25}$$

$$\mathbf{R} = E[\mathbf{C}_n] = \sum_{l \neq J} \mathbf{C}_{xl}^p. \tag{26}$$

When (23) or (24) is employed, (22) becomes an approximation of the MV beamformer. When (25) or (26) is employed, (22) becomes an approximation of the ML beamformer. The choice of the covariance $\mathbf{R}$ is discussed in Section 5.1.

The beamformer is updated every N frames based on the updates of these intermediate variables. A unit consisting of N frames is referred to as "block" hereafter. The relation of frame and block is depicted in Fig.2. Usually, the movement of sources is much slower than the dynamical change of source signal such as speech. When the movement of sources in a block is sufficiently small, it is expected that the signals in this block can be separated by a linear fixed beamformer that is optimized for this block.

4.2. System

In this section, the EM-based sound localization proposed by [1, 2] and the sound separation proposed in the previous section is combined, and the entire procedure is described.

A data unit consisting of N_b blocks is defined as "segment" (see Fig.2.) For obtaining the initial value of the EM-based sound localization, the rough location of sound sources is estimated first by a conventional sound localizer such as MUSIC [4] using all of the data in a segment. Since all of the data in a segment is used, the estimated location is the averaged location when the sound sources move. This estimated location then becomes the initial value of the EM-based sound localization.

In the first block of the segment, the EM algorithm is iterated with the initial values estimated above until $p = P_{max}$.

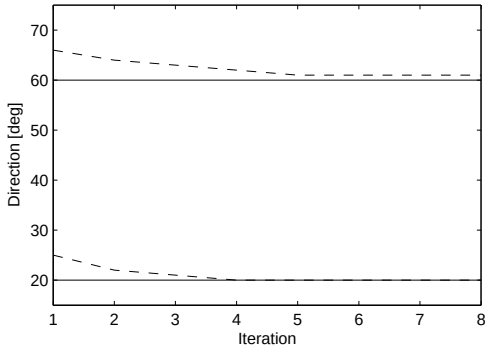

Figure 4: Variation of the estimated location during the iteration of the EM algorithm. Solid line: true location, dashed line: estimate.

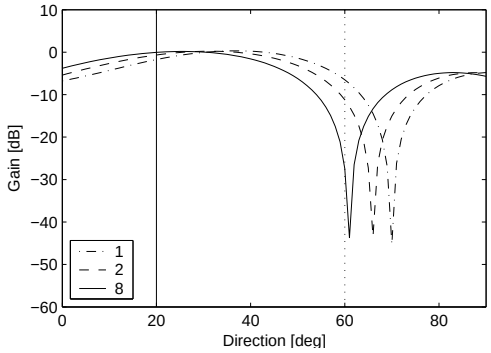

Figure 5: Directivity of the gain matrix $\mathbf{G}_1$.

The final estimate in the iteration $\hat{\Theta}^{P_{max}}$ is employed as the estimated location in this block. Also, the final intermittent variables, $\mathbf{C}_{xl}^{P_{max}}$ or $\hat{\mathbf{K}}_{xl}^{P_{max}}$, is extracted. The beamformer (22) is then constructed and the signals within the block is processed with this beamformer. In the next block, the same procedure is iterated with the final estimates $\hat{\Theta}^{P_{max}}$ and $\hat{\mathbf{K}}_s^{P_{max}}$ in the previous block as the initial values.

5. Simulation

5.1. Static case

Before applying the proposed method to the case with moving sound sources, basic properties of this method are investigated using the case when the location of sources is fixed. In this sub-section, the direction of the sound sources were $\Theta = [20, 60]$. The initial value of the EM algorithm was $[30, 70]$. The input vector was generated using (1) with the theoretical value of $\mathbf{A}$ and the complex random Gaussian noise as $\mathbf{S}(t)$ and $\mathbf{n}(t)$. The case for 1000 Hz was treated. The microphone array used has a circular configuration with a diameter of 0.5 m and consists of 8 microphones.

Figure 3 shows the estimation error defined as $E = (1/L)\sum_l |\hat{\theta}_l - \theta_l|$ as a function of the number of averages N in estimating $\mathbf{C}_y$. The estimation error is calculated as an average of 128 trials. From this, it can be seen that the esti-

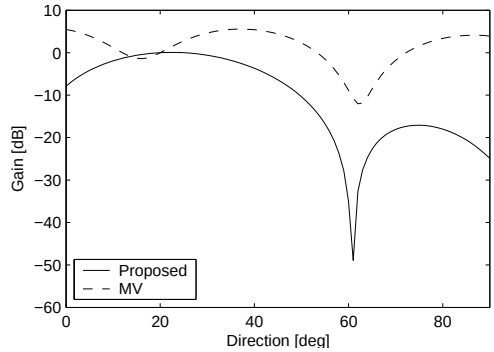

Figure 6: Directivity of the proposed method and the conventional MV beamformer.

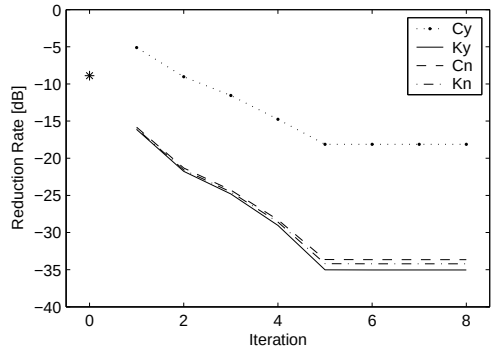

Figure 7: Noise reduction gain of the proposed method as a function of iteration p. The mark $*$ corresponds to that for the conventional MV beamformer.

mation error was reduced to around 1.5° when $N = 8$. For larger N, the estimation error was further reduced. However, for larger N, the tractability for moving sources is reduced. Therefore, $N = 8$ is employed hereafter.

Figure 4 shows the estimated location $\hat{\theta}_l^p$ as a function of iteration p. From this figure, it can be seen that the estimates approach the true values as the iteration proceeds.

Figure 5 shows the directivity of the gain matrix $\mathbf{G}_1$. The directivity was calculated as $\mathbf{P}(\theta) = \mathbf{G}_1\mathbf{a}(\theta)$ where $\mathbf{a}(\theta)$ is the location vector for the arbitrary direction θ, and then the first element was extracted from $\mathbf{P}(\theta)$ which corresponds to the gain for the first microphone. It can be seen that a notch was made in the direction around the source #2 (60°). This is the effect of the gain matrix $\mathbf{G}_1$ as the MV beamformer as described in Section 3.2. It can also be seen that the center of the notch approaches the true direction of source #2 as iteration proceeds. This is the effect of the model covariance $\hat{\mathbf{K}}_y$ being improved by the iteration of the EM algorithm.

Figure 6 shows the directivity of the proposed beamformer (22) with $\mathbf{R} = \hat{\mathbf{K}}_y$. For the sake of comparison, the directivity of the conventional MV beamformer with the sample covariance $\mathbf{C}_y$ is also shown. It can be seen that while the attenuation for the noise source (60°) is around 10 dB for the conventional MV beamformer, a deep notch was made in 60° for the proposed method. This is the effect of cross-terms

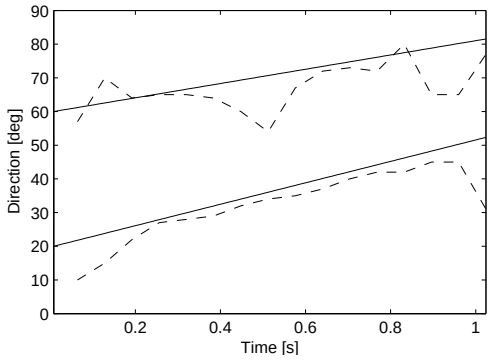

Figure 8: Estimated (dashed line) and true (solid line) trajectory of sources.

not being included in the covariance for the proposed method as described in Section 3.2.

Figure 7 shows the noise reduction gain defined as a ratio of the gain of the beamformer in the direction of the target source (source #1) and that in the direction of the noise source (source #2). The results for four options for $\mathbf{R}$ indicated in Eqs.(23)-(26) are shown. A similar performance was obtained for $\mathbf{R}$ =$\mathbf{K}_y, E[\mathbf{C}_n], \mathbf{K}_n$, while the performance is reduced for $\mathbf{R} = E[\mathbf{C}_y]$. Hereafter, $\mathbf{R}$ =$\mathbf{K}_y$ which showed the best performance in Fig.7 is employed as a covariance for the beamformer (22).

5.2. Dynamic case

In this section, the proposed method is applied to moving sources. For simulating moving sources, matrix $\mathbf{A}$ was dynamically changed. The initial position of the two sources was $[20, 60]^\circ$. The sources moved on the circle with a diameter of 1.5 m with a velocity of $[3, 2]$ km/s. The trajectory of the sources is shown by the solid line in Fig.8. The level of noise $\mathbf{n}(t)$ was -20 dB relative to that of the signal $\mathbf{A}\mathbf{s}(t)$. The frequency component at 1000 Hz of the speech signal was employed as a source signal $\mathbf{s}(t)$.

Figure 8 shows the estimated trajectory (dashed line) for the two sources. From this, it can be seen that the true trajectory is well estimated. However, when the signal for source #2 is weak in the period [0.4,0.6] s (see Fig.9), the estimated trajectory differed from the true trajectory.

Figure 10 shows the spectral component separated by the proposed method using the beamformer. From this, it can be seen that the original waveform (Fig.9) is well recovered by the proposed method. This is due to the fact that the moving velocity of the sources was slow (walking speed) and the movement of the sources in a block (64 ms) was sufficiently small.

6. Conclusion

In this paper, a method of sound source tracking and separation based on the EM algorithm was proposed. The results of simulation confirmed its applicability to the separa-

tion of signals from moving sources. However, this study is in its early stage and the simulation conducted in this paper was very simple and was not realistic. In the next stage, the proposed method should be extended to the broad-band case and should be applied to a problem in a real environment. Also, combining this method with tracking techniques such as Kalman filter [5] or particle filter [8] would further enhance the separation performance.

7. Appendix: Derivation of the expected value of the covariance

In the linear measurement process given by $\mathbf{y} = \mathbf{H}\mathbf{x}$ where $\mathbf{x}$ is a 0-mean complex Gaussian random vector, the least square estimate of $\mathbf{x}$ and its error covariance $\mathbf{P} = \mathrm{Cov}[\mathbf{x}-\hat{\mathbf{x}}]$ is given as [7]

$$\hat{\mathbf{x}} = \mathbf{K}_x\mathbf{H}^H(\mathbf{H}\mathbf{K}_x\mathbf{H}^H)^{-1}\mathbf{y} \quad (27)$$

$$\mathbf{P} = \mathbf{K}_x - \mathbf{K}_x\mathbf{H}^H(\mathbf{H}\mathbf{K}_x\mathbf{H}^H)^{-1}\mathbf{H}\mathbf{K}_x \quad (28)$$

where $\mathbf{K}_x = E[\mathbf{x}\mathbf{x}^H]$. Using $\mathbf{K}_y = \mathbf{H}\mathbf{K}_x\mathbf{H}^H$ and defining $\mathbf{G}$ as

$$\mathbf{G} = \mathbf{K}_x\mathbf{H}^H(\mathbf{H}\mathbf{K}_x\mathbf{H}^H)^{-1} = \mathbf{K}_x\mathbf{H}^H\mathbf{K}_y^{-1}, \quad (29)$$

(27) and (28) are rewritten as

$$\hat{\mathbf{x}} = \mathbf{G}\mathbf{y} \quad (30)$$

$$\mathbf{P} = \mathbf{K}_x - \mathbf{G}\mathbf{H}\mathbf{K}_x \quad (31)$$

On the other hand, since $\mathbf{x}$ is a Gaussian random vector, the least square estimate, $\hat{\mathbf{x}}$, also becomes the maximum likelihood estimator:

$$\hat{\mathbf{x}} = \int \mathbf{x}p(\mathbf{x}|\mathbf{y})d\mathbf{x} \quad (32)$$

$$\mathbf{P} = \int (\mathbf{x}-\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})^H p(\mathbf{x}|\mathbf{y})d\mathbf{x} \quad (33)$$

where $p(\mathbf{x}|\mathbf{y})$ denotes the *a posteriori* density. Using (32) and (33), the conditional expectation of $\mathbf{C}_x$, given $\mathbf{y}$, is

$$E[\mathbf{C}_x|\mathbf{y}] \equiv \int \mathbf{x}\mathbf{x}^H p(\mathbf{x}|\mathbf{y})d\mathbf{x} = \mathbf{P} + \hat{\mathbf{x}}\hat{\mathbf{x}}^H \quad (34)$$

Substituting (30) and (31) into (34) and defining $\mathbf{C}_y = \mathbf{y}\mathbf{y}^H$,

$$E[\mathbf{C}_x|\mathbf{y}] = \mathbf{K}_x - \mathbf{G}\mathbf{H}\mathbf{K}_x + \mathbf{G}\mathbf{C}_y\mathbf{G}^H \quad (35)$$

Substituting (29) into (35),

$$E[\mathbf{C}_x|\mathbf{y}] = \mathbf{K}_x - \mathbf{K}_x\mathbf{H}^H\mathbf{K}_y^{-1}\mathbf{H}\mathbf{K}_x + \mathbf{K}_x\mathbf{H}^H\mathbf{K}_y^{-1}\mathbf{C}_y\mathbf{K}_y^{-H}\mathbf{H}\mathbf{K}_x^H \quad (36)$$

Let us define $\mathbf{K}_{x1}, \cdots, \mathbf{K}_{xM}$ as

$$[\mathbf{K}_{x1} \cdots \mathbf{K}_{xM}] = \mathbf{H}\mathbf{K}_x \quad (37)$$

Extracting the lth block from (36) using (37), and substituting $\mathbf{C}_y = (1/N)\sum_{t=1}^{N}\mathbf{y}(t)\mathbf{y}(t)^H$ for $\mathbf{C}_y = \mathbf{y}\mathbf{y}^H$, we obtain (6).

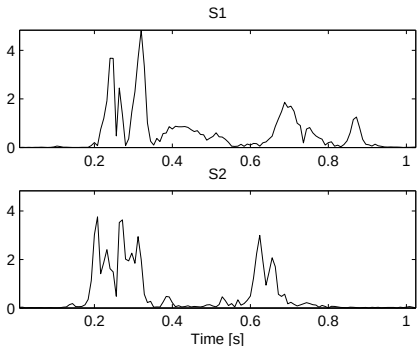

Figure 9: Source signal (frequency component at 1000 Hz).

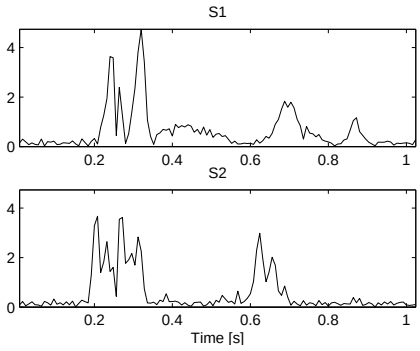

Figure 10: Signals separated by the proposed method.

8. References

[1] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 36, no. 4, pp. 477–489, 1988.

[2] M. Miller and D. Fuhrmann, "Maximum-likelihood narrowband direction finding and the EM algorithm," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 38, no. 9, pp. 1560–1577, 1990.

[3] F. Asano, *et al.*, "Detection and separation of speech event using audio and video information fusion and its application to robust speech interface," *accepted for J. Applied Signal Processing*, 2004.

[4] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag*, vol. AP-34, no. 3, pp. 276–280, March 1986.

[5] R. Zarnich, K. Bell, and H. VanTrees, "A unified method for measurement and tracking of contacts from an array of sensors," *IEEE Trans. Signal Process*, vol. 49, no. 12, pp. 2950–2961, Dec. 2001.

[6] Don H. Johnson and Dan E. Dudgeon, *Array signal processing*, Prentice Hall, Englewood Cliffs NJ, 1993.

[7] S. Arimoto, *Kalman Filter (in Japanese)*, Sangyo-Tosho, Tokyo,Japan, 1977.

[8] H. Asoh, *et al.*, "An application of a particle filter to bayesina multiple sound source tracking with audio and video information fusion," *Proc. Fusion 2004*, 2004