Stereo Source Separation and Source Counting with MAP Estimation with Dirichlet Prior Considering Spatial Aliasing Problem

Shoko Araki, Tomohiro Nakatani, Hiroshi Sawada, and Shoji Makino

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{shoko,nak,sawada,maki}@cslab.kecl.ntt.co.jp

Abstract. In this paper, we propose a novel sparse source separation method that can estimate the number of sources and time-frequency masks simultaneously, even when the spatial aliasing problem exists. Recently, many sparse source separation approaches with time-frequency masks have been proposed. However, most of these approaches require information on the number of sources in advance. In our proposed method, we model the phase difference of arrival (PDOA) between microphones with a Gaussian mixture model (GMM) with a Dirichlet prior. Then we estimate the model parameters by using the maximum a posteriori (MAP) estimation based on the EM algorithm. In order to avoid one cluster being modeled by two or more Gaussians, we utilize a sparse distribution modeled by the Dirichlet distributions as the prior of the GMM mixture weight. Moreover, to handle wide microphone spacing cases where the spatial aliasing problem occurs, the indeterminacy of modulus $2\pi k$ in the phase is also included in our model. Experimental results show good performance of our proposed method.

Keywords: Dirichlet distribution, prior, number of sources, blind source separation, sparse, spatial aliasing problem.

1 Introduction

Blind source separation (BSS) is an approach for estimating source signals that uses only the mixed signal information observed at each microphone. The BSS technique for speech dealt with in this paper has many applications, including the hands-free teleconference systems and preprocessing for an automatic speech recognizer.

Let us formulate the task. Suppose that $N_s \geq 2$ speech sources $\mathrm{s}_1, \ldots, \mathrm{s}_{N_s}$ are convolutively mixed and observed at N_m microphones,

$$\mathrm{x}_j(t) = \sum_{i=1}^{N_s} \sum_l \mathrm{h}_{ji}(l)\, \mathrm{s}_i(t-l),\ j=1,\ldots,N_m, \tag{1}$$

where $\mathrm{h}_{ji}(l)$ represents the impulse response from source i to microphone j. Our goal is to obtain estimates y_i of each source signal s_i from the microphone observations x_j without information about the number of sources N_s, the speech sources s_i or the mixing process h_{ji}.

T. Adali et al. (Eds.): ICA 2009, LNCS 5441, pp. 742–750, 2009.

Two approaches have been widely studied and employed to solve the BSS problem: one is based on independent component analysis (ICA) (e.g., [1]) and the other relies on the sparseness of source signals (e.g., [2]). In this paper, we focus on the latter approach, more specifically, the time-frequency mask approach [2,3]. With the time-frequency mask approach, we classify the phase difference of arrival (PDOA) between microphone observations, and separate each signal by collecting the observation signal at time-frequency points in each cluster.

In previous work [2,3], to automatically find clusters, the number of sources N_s is assumed to be known. However, in real situations we usually cannot obtain information on the number of sources N_s in advance. Especially for an underdetermined case ($N_s > N_m$), the source counting is difficult, and few papers have dealt with this problem. Moreover, when the microphone spacing is large, the spatial aliasing problem occurs. This problem makes it difficult to classify the PDOA because the phase has the indeterminacy of modulus $2\pi k$ in high frequencies. [4] considered the spatial aliasing problem in a time-frequency mask approach, however, the number of sources N_s should be known.

In this paper, we propose a novel sparse source separation method that can estimate the number of sources and time-frequency masks simultaneously, even when spatial aliasing occurs. We model the PDOA distribution with a Gaussian mixture model (GMM) with a Dirichlet prior [5], and estimate the model parameters by using the EM algorithm. In order to avoid one cluster being modeled by two or more Gaussians, thus making it possible to estimate the number of sources correctly, we propose utilizing a sparse distribution modeled by the Dirichlet distribution as the prior of the GMM mixture weight. The authors of [6,7] also derived the EM algorithm, however, they still needed to know the number of sources N_s in advance. On the other hand, our proposed algorithm does not require information on the source number, thanks to the weight prior. Because the indeterminacy of $2\pi k$ in phase is modeled in our GMM, we can also overcome the difficulty in the PDOA clustering even in a spatial aliasing case.

The experimental results with a wide microphone spacing (20 cm) show that our proposed method can estimate the number of sources and can separate signals by time-frequency masks obtained by the posterior probability for each cluster.

2 Mixing and Separation Processes

This paper employs a time-frequency domain approach. With an F-point short-time Fourier transform (STFT), (1) is converted into:

$$x_j(n,f) = \sum_{i=1}^{N_s} h_{ji}(f) s_i(n,f), \tag{2}$$

where $h_{ji}(f)$ is the frequency response from source i to microphone j, $s_i(n,f)$ is the STFT of a source s_i. $f \in \{0, \frac{1}{F}f_s, \ldots, \frac{F-1}{F}f_s\}$ is a frequency (f_s is the sampling frequency) and $n(=0,\cdots,N-1)$ is a time-frame index.

In this paper, we assume the sparseness of the sources [2]:

$$x_j(n,f) \approx h_{ji}(f) s_i(n,f), \tag{3}$$

where $s_i(n,f)$ is a dominant source at the time-frequency slot (n,f). This is approximately true for speech signals in the time-frequency domain [2,3].

2.1 Separation Method

In this paper, we assume that $h_{ji}(f)$ in (2) is modeled by an anechoic model (e.g., eq. (13) of [3]), that is, the PDOA between microphones is given as:

$$\arg\left[\frac{x_1(n,f)}{x_2(n,f)}\right] = 2\pi f\tau(n,f) = 2\pi f\frac{d\cos\varphi(n,f)}{v}, \tag{4}$$

where $\tau(n,f) = d\cos\varphi(n,f)/v$ is the time difference of arrival (TDOA), $\varphi(n,f)$ is the dominant source direction at the time-frequency (n,f), and d and v denote the microphone spacing and the sound speed.

First, by assuming the source sparseness, we calculate the PDOA at each time-frequency slot by the left-side of (4). Then, by considering the frequency dependence in the PDOA, we classify the PDOA values in some way. For example, if there is no spatial aliasing problem and we know the number of sources, the k-means clustering algorithm can be applied to TDOA $\tau(n,f) = \frac{1}{2\pi f}\arg[x_1(n,f)/x_2(n,f)]$. Our method, which considers the aliasing problem and the unknown source number, is introduced in the following section.

Finally, we estimate the separated signals $y_i(n,f)$ with time-frequency masks $M_i(n,f)$, which extract time-frequency points of members in the i-th cluster:

$$y_i(n,f) = x_1(n,f)M_i(n,f). \tag{5}$$

3 Proposed Method

3.1 Problems in PDOA Clustering

The first problem for the PDOA clustering is the spatial aliasing problem. As can be seen in (4), when the frequency f or the microphone spacing d are large, $\arg[x_1(n,f)/x_2(n,f)] = 2\pi f d\cos\varphi(n,f)/v$ exceeds $\pm\pi$. However, since the arg operation has the indefiniteness of modulus $2\pi k$, (4) should be:

$$2\pi f\tau(n,f) = \arg\left[\frac{x_1(n,f)}{x_2(n,f)}\right] + 2\pi k = o(n,f) + 2\pi k, \tag{6}$$

where $o(n,f) = \arg[x_1(n,f)/x_2(n,f)]$, $-\pi \le o(n,f) < \pi$, and k is an integer. Note that we can observe just $o(n,f)$, and k is unknown when the source direction φ is unknown. This is the spatial aliasing problem. Figure 1 gives an example of the observed PDOA $o(n,f)$ for a wide microphone spacing of 20 cm. In the next subsection, k in (6) is considered as a hidden variable.

The second problem occurs when we apply a GMM fitting method for an unknown number of mixtures. Figure 2 shows an example. Here we have two clusters in the histogram (Fig. 2(a)). Figure 2(b) shows the fitting result of GMM of eight Gaussians, to Fig. 2 (a). From Fig. 2(b), we can see that multiple Gaussians are fit to each cluster. However, we expect just one Gaussian for each peak, in order to estimate the number of sources by counting the number of dominant Gaussians. In this paper, in order to avoid the case where one cluster is modeled by two or more Gaussians, we propose utilizing a sparse distribution for the prior of the GMM mixture weight parameter in the next subsection.

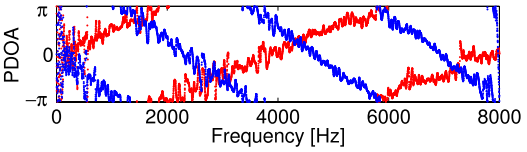

Fig. 1. Example PDOA for a microphone spacing of 20 cm and a sampling rate of 16 kHz. The PDOA for a source at 70° and a source at 150° are drawn individually for illustrative purposes.

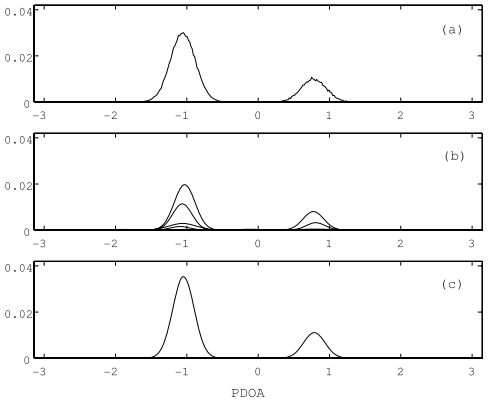

Fig. 2. Example GMM fitting result with and without prior. (a) Histogram of two Gaussians, (b) estimated Gaussians of GMM without prior ($\phi = 1.0$), (c) estimated Gaussians of GMM with prior ($\phi = 0.9$).

3.2 Probabilistic Model

To begin with, let us consider that we observe one source from one direction. Hereafter, a notation $o_{nf} = o(n, f)$ is utilized. Because the spatial aliasing issue in (6) can be considered as a phase wrapping problem, we can model the PDOA with a Gaussian distribution by considering the unwrapped data $o_{nf} + 2\pi k$. In other words, the phase wrapping process can be modeled by summing the Gaussians at intervals of 2π. That is, we assume that the PDOA follows a wrapped Gaussian distribution [8],

$$p(o_{nf}; \mu, \sigma) = \sum_{k=-K_f}^{K_f} p(o_{nf}, k; \mu, \sigma) = \sum_{k=-K_f}^{K_f} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(o_{nf} + 2\pi k - 2\pi f\mu)^2}{2\sigma^2}\right), \tag{7}$$

where $-\pi \leq o_{nf} < \pi$, μ gives us the expectation value of the TDOA τ of the source, σ^2 is the variance of the PDOA, and k is an integer to handle the spatial aliasing (6). The value K_f is a frequency dependent integer, and it can be determined if we know the microphone spacing d and the frequency f. If we do not know d, we can set a sufficiently large value for K_f for all frequencies. This model is inspired by a wrapped Gaussian model [8].

In our observed mixture, we assume that there are a sufficient number of source signals from different directions, where some are dominant and others are much less dominant. Each source is modeled by (7). We also assume that the PDOA for an observed mixture follows a Gaussian mixture model (GMM):

$$p(o_{nf};\mu_m,\sigma_m)=\sum_{m=1}^{M}\sum_{k=-K_f}^{K_f}\frac{\alpha_m}{\sqrt{2\pi\sigma_m^2}}\exp\left(\frac{-(o_{nf}+2\pi k-2\pi f\mu_m)^2}{2\sigma_m^2}\right). \quad (8)$$

We prepare a sufficient number M of Gaussians for our GMM model and estimate the mean μ_m, variance σ_m^2 and weight α_m for each Gaussian m.

In order to solve the second problem mentioned in Section 3.1, that is, in order to model the observed PDOA data by allocating one Gaussian to each source, we assume *the sparseness of the source directions*, where each direction is dominated by at most one source. For this purpose, as the prior of the mixture weight, we employ the Dirichlet distribution:

$$p(\alpha)=\frac{1}{B(\phi)}\prod_{m}^{M}\alpha_m^{\phi-1}, \quad (9)$$

where $\alpha=\{\alpha_1,\cdots,\alpha_m,\cdots,\alpha_M\}$, $\sum_m^M\alpha_m=1$, $0\le\alpha_m\le 1$, and $B(\phi)$ is the beta distribution (regularization term). When we set small hyper parameter ϕ ($\phi<1$), the prior takes a larger value as the number of mixture weights whose values are close to zero increases, which is desirable for representing the sparseness of the source direction [5]. In addition, the Dirichlet distribution is known to be a conjugate prior of the mixture weight [5], and it can be incorporated into the GMM fitting approach in a computationally efficient manner.

Figure 2 (c) shows a GMM fitting result with prior ((9) with $\phi=0.9$) for the distribution in Fig. 2 (a). In spite of utilizing eight Gaussians, we can see that just two Gaussians are dominant in Fig. 2 (c). That is, using the prior, more correct GMM fitting can be performed.

3.3 Cost Function Based on GMM

Let $\theta=\{\alpha_m,\mu_m,\sigma_m\}$ be a model parameter set. The observations are $o=\{o_{11},o_{12},\cdots,o_{nf},\cdots,o_{NF}\}$ and power values $a=\{a_{11},a_{12},\cdots,a_{nf},\cdots,a_{NF}\}$, where $a_{nf}=a(n,f)=|x_1(n,f)|^2$. In the following, Gaussian indices m and k in the PDOA model (8) are assumed not to be observed, and therefore dealt with as hidden variables.

The cost function of the maximum a posteriori (MAP) estimation is defined based on a log of a joint probability density function (pdf) as

$$\mathcal{L}(\theta)=\log p(o,\theta)=\log p(o|\theta)+\log p(\alpha)+\text{const.} \quad (10)$$

$$=\sum_{n}^{N}\sum_{f}^{F}f(a_{nf})\log p(o_{nf}|\theta)+\log p(\alpha)+\text{const.} \quad (11)$$

$$=\sum_{n}^{N}\sum_{f}^{F}f(a_{nf})\log\left(\sum_{m}^{M}\sum_{k=-K_f}^{K_f}p(m,k,o_{nf}|\theta)\right)+\log p(\alpha)+\text{const.}, \quad (12)$$

where

$$p(m,k,o_{nf}|\theta) = \frac{\alpha_m}{\sqrt{2\pi\sigma_m^2}} \exp\left(\frac{-(o_{nf}+2\pi k-2\pi f\mu_m)^2}{2\sigma_m^2}\right). \quad (13)$$

We disregarded the priors of the model parameters except for α in (10), and

$$f(a_{nf}) = ca_{nf}/\sum_n\sum_f a_{nf} \quad (14)$$

gives a power weight ($a_{nf} = |x_1(n,f)|^2$ in this paper) and controls the importance of the observation relative to the prior term (2nd term of (12)), where c is a control parameter.

In (12), the mixture weight α follows the Dirichlet distribution (9) and $\sum_m^M \alpha_m = 1$, $0 \le \alpha_m \le 1$ holds. For the sparse representation of the GMM, $\phi < 1$ is preferred for the Dirichlet distribution (9). Note that $\phi = 1$ is equivalent to the case without a prior for the mixture weight.

3.4 EM Algorithm

Here we derive an algorithm for estimating parameter θ by the EM algorithm.

The auxiliary function Q is given as

$$Q(\theta|\theta^t) = \mathrm{E}\left[\log p(o_{nf};\theta)|o_{nf};\theta^t\right] \quad (15)$$

$$= \sum_n\sum_f\sum_m\sum_k \left[p(m,k|o_{nf},\theta^t)f(a_{nf})\log p(m,k,o_{nf}|\theta)\right] + \log p(\alpha), \quad (16)$$

where θ^t is the estimate of the parameters after the t-th iteration, and

$$p(m,k|o_{nf},\theta^t) = \frac{p(m,k,o_{nf}|\theta^t)}{\sum_m\sum_k p(m,k,o_{nf}|\theta^t)}. \quad (17)$$

By setting $\frac{\partial Q(\theta|\theta^t)}{\partial \mu_m} = 0$ and $\frac{\partial Q(\theta|\theta^t)}{\partial \sigma_m^2} = 0$, we obtain

$$\mu_m^{t+1} = \frac{\sum_n\sum_f\sum_k p(m,k|o_{nf},\theta^t)f(a_{nf})(o_{nf}+2\pi k)}{\sum_n\sum_f\sum_k 2\pi f p(m,k|o_{nf},\theta^t)f(a_{nf})} \quad (18)$$

$$(\sigma_m^2)^{t+1} = \frac{\sum_n\sum_f\sum_k p(m,k|o_{nf},\theta^t)f(a_{nf})(o_{nf}+2\pi k-2\pi f\mu_m)^2}{\sum_n\sum_f\sum_k p(m,k|o_{nf},\theta^t)f(a_{nf})}. \quad (19)$$

Moreover, by using the Lagrange multiplier method, $\sum_m^M \alpha_m = 1$ and (14), the mixture weight is obtained as follows:

$$\alpha_m^{t+1} = \frac{1}{c+M(\phi-1)}\left\{\sum_n\sum_f\sum_k p(m,k|o_{nf},\theta^t)f(a_{nf}) + (\phi-1)\right\}. \quad (20)$$

Since $\alpha_m > 0$, $c > M(1-\phi)$ must hold from (20).

In the E-step we calculate (17), then in the M-step the parameters θ are calculated by using (18), (19) and (20). Sometimes $\alpha_m < 0$ occurs. In such a case, we can factor out the corresponding Gaussian (by setting $\alpha_m = \epsilon$, where ϵ is a very small value) and recalculate the parameters.

3.5 Source Counting

Thanks to the Dirichlet prior (9), most of the mixture weight parameter α_m becomes very close to zero and some have dominant values. Sometimes, some weight parameters α_m do not come to zero sufficiently because of very large variance σ_m. So, we can determine the number of sources N_s by counting the number of Gaussians whose parameters meet conditions $\alpha_m \geq \epsilon$ and $\sigma_m \leq th$, where ϵ is a sufficiently small threshod value and th is an appropriate threshold value. $\epsilon = 0.2$ and $th = \pi/3$ degrees are used in this paper.

3.6 Source Separation

The time-frequency mask $M_m(n, f)$ for the m-th separated source (see (5)) is obtained by marginalizing the estimated pdf (17) with respect to k,

$$M_m(n, f) = p(m|o_{nf}, \theta) = \sum_{k=-K_f}^{K_f} p(m, k|o_{nf}, \theta). \tag{21}$$

The separated signal is obtained by

$$y_m(n, f) = x_1(n, f) M_m(n, f) = x_1(n, f) p(m|o_{nf}, \theta). \tag{22}$$

4 Experiments

4.1 Experimental Setup

We performed experiments with measured impulse responses h_{ji} in a room whose reverberation time was 130 ms (see Fig. 9's setup A of [4]). We utilized two microphones whose spacing was 20 cm. The numbers of sources N_s were two and three. Mixtures were made by convolving the measured room impulse responses and 5-second English speech signals sampled at 16 kHz. The frame size F for STFT was 1024 (64 ms), and the frame shift was 256 (16 ms).

In the EM algorithm, we utilized $M = 8$ Gaussians. From the microphone spacing and sampling rate, the aliasing problem occured above 850 Hz. In our implementation, $K_f = K = 5$ was utilized for all frequencies f. For the comparison with an aliasing-unconsidered case, $K_f = K = 0$ for all frequencies f was also tested. As the hyper parameter for (9), we utilized $\phi = 0.9$ for our proposed method and $\phi = 1.0$ for a conventional EM algorithm that corresponds to the case without any prior for the mixture weights. The number of iterations was 10, and the control parameter c for (14) was 5.

We evaluated the signal-to-interference ratio (SIR) as a separation performance measure, and the signal-to-distortion ratio (SDR) as a sound quality measure. Their definitions can be found in [3]. We calculated SIR and SDR values for the separated sources that are counted as the sources by the method in Section 3.5. We conducted 20 trials with different speech source combinations and location combinations, and then averaged the results.

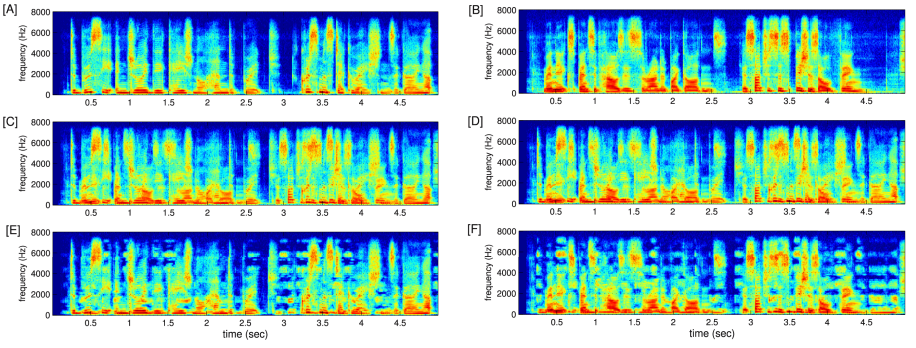

Fig. 3. Example spectra of (A)(B) sources, (C)(D) observations, (E)(F) separated signals. $N_s = 2$, $\phi = 0.9$ and $K = 5$.

Table 1. Experimental results. Input SIR was 0.0 [dB] ($N_s = 2$), and -3.1 [dB] ($N_s = 3$).

N_s			Accuracy of $\hat{N}_s$ estimation [%]								Performance [dB]	
	ϕ	K	$\hat{N}_s$:1	2	3	4	5	6	7	8	Output SIR	SDR
2	0.9	5		100							11.5	12.4
	0.9	0		35	60	5					11.4	7.5
	1.0	5		0	5	40	35	20			8.8	9.7
3	0.9	5		15	75	10					7.5	8.2
	0.9	0	5	25	50	20					2.0	8.7
	1.0	5			0	90	10				6.9	7.5

4.2 Results

Figure 3 shows the example spectra of sources, observations, and separated sources for a two source case. The source directions were 70° and 150°, whose example PDOA is shown in Fig. 1. By comparing the source spectra Fig. 3 (A)(B) and the separation spectra Fig. 3 (E)(F), it can be seen that the spatial aliasing problem does not occur in most frequencies. However, it is also seen that at the frequencies where the PDOA of two sources lap over each other, say around 1500, 3000, 4500, 6000, 7500 Hz (see Fig. 1), the signals are not separated well. Such phenomena can be seen in the separated spectra Fig. 3 (E)(F).

Table 1 reports the experimental results. In the table, $\phi = 0.9$ means the results with sparse prior (9) and $\phi = 1.0$ indicates the results without a prior. $K = 5$ and $K = 0$ mean the spatial aliasing is considered and unconsidered, respectively. The percentage values are shown where the method estimates the number of sources as $\hat{N}_s$ within 20 trials. The average separation performance results, SIR and SDR in dB, are also reported.

From Table 1, we can see that with the prior ($\phi = 0.9$) by considering the aliasing ($K = 5$), the number of sources is almost perfectly estimated. On the other hand, without the prior, the number of sources is overestimated, and the accuracy rate was quite low.

As for the separation performance, we obtained better performance by using the prior ($\phi = 0.9$) than without the prior ($\phi = 1.0$) when $K = 5$. When we did not consider the spatial aliasing, $K = 0$, the separation performance was of course poor, especially when $N_s = 3$.

5 Conclusion

We proposed a speech source separation method that can estimate both the number of sources and separation masks. We model the PDOA with a GMM, where the phase indefiniteness in spatial aliasing cases is considered. We employ the Dirichlet distribution as the prior of the GMM mixture weight to model each cluster by a single Gaussian. Our experimental results show that the proposed method can estimate the number of sources correctly. We also confirmed that the proposed method gives good separation performance in a room with reverberation time of 130 ms.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, Chichester (2001)
2. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Processing 52(7), 1830–1847 (2004)
3. Araki, S., Sawada, H., Mukai, R., Makino, S.: Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. Signal Processing 77(8), 1833–1847 (2007)
4. Sawada, H., Araki, S., Mukai, R., Makino, S.: Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. IEEE Trans. Audio, Speech and Language Processing 15(5), 1592–1604 (2007)
5. Bishop, C.M.: Pattern recognition and machine learning. Springer, Heidelberg (2008)
6. Mandel, M., Ellis, D., Jebara, T.: An EM algorithm for localizing multiple sound sources in reverberant environments. In: Proc. Neural Info. Proc. Sys. (2006)
7. O'Grady, P., Pearlmutter, B.: Soft-LOST: EM on a mixture of oriented lines. In: Puntonet, C.G., Prieto, A.G. (eds.) ICA 2004. LNCS, vol. 3195, pp. 430–436. Springer, Heidelberg (2004)
8. Smaragdis, P., Boufounos, P.: Learning source trajectories using wrapped-phase hidden markov models. In: Proc. of WASPAA 2005, pp. 114–117 (2005)