On Initial Seed Selection for Frequency Domain Blind Speech Separation

Dang Hai Tran Vu and Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany

{tran,haeb}@nt.uni-paderborn.de

Abstract

In this paper we address the problem of initial seed selection for frequency domain iterative blind speech separation (BSS) algorithms. The derivation of the seeding algorithm is guided by the goal to select samples which are likely to be caused by source activity and not by noise and at the same time originate from different sources. The proposed algorithm has moderate computational complexity and finds better seed values than alternative schemes, as is demonstrated by experiments on the database of the SiSEC2010 challenge.

Index Terms: Blind source separation, seed selection, initialization, expectation maximization, independent component analysis

1. Introduction

Techniques for the blind separation of simultaneously active speech sources have drawn increased attention in the speech processing community in the last decade motivated by the desire to realize telecommunication scenarios with distant-talking microphones. To meet the challenges of the degraded signal quality compared to the use of close-talking microphones, multichannel processing is preferred due to the performance gains by exploiting spatial information.

Two approaches have been mainly studied in the last years: sparseness-based blind source separation (SBSS) and complex-valued independent component analysis (CVICA). In SBSS the working assumption is that in each time-frequency slot at most one source is active at a time. The goal is then to determine the active source and its corresponding transmission characteristics to the microphones. Recently various methods based on the expectation maximization (EM) framework have been developed to uncover source activity and to identify the mixing matrix at the same time, e.g., [1, 2, 3, 4, 5]. CVICA techniques minimize the statistical dependency of the estimated sources, e.g., [6, 7]. These methods usually involve non-linear contrast functions which are derived from the probability density function of the frequency-domain source signal representation.

Both SBSS and CVICA are usually implemented by iterative algorithms such as the method of steepest descent or the aforementioned EM algorithm where, starting with an initial guess, parameters are updated by local optimization strategies.

In this contribution we are concerned with the selection of the initial seed values for the iterative algorithms. This is an important issue as it is well known that inappropriate seed selection may result in slow convergence or current estimates being stuck in local optima. While seed value selection is an issue for both SBSS and CVICA, we will concentrate on SBSS in this paper. Here, seed values are needed for the columns of the mixing matrix $\mathbf{H}$, as they describe the transmission characteristics from the sources to the sensors, see section 2 below.

One of the simplest method is to draw uniformly at random I observation vectors, where I is the number of sources, and initialize the columns of the mixing matrix with these observations after normalization to unit length [5]. While being computationally inexpensive, this technique makes no effort either to obtain seeds close to the optimum, i.e. the true values of the columns of the mixing matrix $\mathbf{H}$, or to represent each of the sources in the seed values. As a result, the seed values can be arbitrarily poor.

Other schemes are based on assumptions or estimates of the direction-of-arrival (DoA). In [7] DoA estimation with null-beamforming and iterative CVICA update are closely merged resulting in an improved convergence of the algorithm. However, DoA-based seed selection is, strictly speaking, not a blind approach since a priori knowledge about the array geometry is required, thus contradicting to original BSS assumptions.

In [3] the authors propose to use an initialization scheme based on hierarchical agglomerative clustering. At the beginning each observation represents a cluster. Then the closest clusters are successively merged until a predetermined number of $\tilde{I}$ clusters is reached, which is much larger than the number of sources I. Finally, the I clusters with the largest membership count are chosen as seeds. While outliers are eliminated in this way, hierarchical clustering is computationally expensive, as it requires the computation of the pairwise distances among the observations.

The key idea of the method presented in the following is the selection of observations guided by two principles. First, the seed should contain most probably speech and not noise, and second, the seed is chosen such, that most probably it is not a representative of an already chosen source. In the following we explain how these principles can be realized in practice.

2. Signal Model

In multichannel blind speech separation we are given recordings $x_1(t), ..., x_D(t)$ of a D-element microphone array, which are mixtures of I filtered speech sources $s_1(t), ..., s_I(t)$:

$$x_d(t) = \sum_{i=1}^{I} \sum_{l} h_{id}(l)\; s_i(t-l) + n_d(t) \quad d = 1, ..., D. \tag{1}$$

Here, $h_{id}(l)$ is the unknown impulse response from source i to sensor d, and $n_1(t), ..., n_D(t)$ denotes the additive noise, which is present at each sensor. The goal of BSS is to recover the source signals s_i, $i = 1, \ldots, I$, solely from the observations x_d, $d = 1, \ldots, D$.

Most of the existing algorithms for BSS transform the signals into the time-frequency domain with the short-time Fourier transform (STFT). After switching to the frequency domain,

This work was in part supported by Deutsche Forschungsgemeinschaft under contract number HA 3455/8-1.

separation can now be carried out in each frequency bin separately. Using vector notation we have

$$\mathbf{X}(m,k) = \mathbf{H}(k)\,\mathbf{S}(m,k) + \mathbf{N}(m,k), \tag{2}$$

where $\mathbf{X} = [X_1, \ldots, X_D]^T$ is the observation vector in the STFT domain, $\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_I]$ is the $D \times I$ mixing matrix consisting of the transfer function vectors $\mathbf{H}_i = [H_{i1}, \ldots, H_{iD}]^T$, $\mathbf{S} = [S_1, \ldots, S_I]^T$ is the source vector and $\mathbf{N} = [N_1, \ldots, N_D]^T$ is the noise vector. Further, m and k are the frame and frequency bin index, respectively.

The BSS algorithm has to find a possibly time-variant demixing matrix $\mathbf{W}$ from which estimates $\hat{\mathbf{S}}$ of the original speech vectors $\mathbf{S}$ can be obtained by

$$\hat{\mathbf{S}}(m,k) := \mathbf{W}(m,k)\,\mathbf{X}(m,k). \tag{3}$$

Before reconstructing the time-domain signals the permutation and scaling indeterminacy has to be solved. However, in this paper we are not concerned with these issues and assume that the columns of $\mathbf{H}$ and $\mathbf{W}$ have unit Euclidean norm.

3. Proposed seed selection method

Seed value selection is concerned with finding initial estimates of $\mathbf{H}$ (or, equivalently, $\mathbf{W}$). The proposed approach is tied to the sparseness assumption, which states that at any time-frequency slot (m,k) at most one source is active. This assumption can be formulated by introducing the hidden random variable $Z(m,k) \in \{0, \ldots, I\}$:

$$Z(m,k) = 0 :$$

$$\mathbf{X}(m,k) = \mathbf{N}(m,k) \tag{4}$$

$$Z(m,k) = i, \qquad i \in \{1, \ldots, I\} :$$

$$\mathbf{X}(m,k) = \mathbf{H}_i(k) S_i(m,k) + \mathbf{N}(m,k), \tag{5}$$

where $Z(m,k) = i$, $i \in \{1, \ldots, I\}$, indicates that source i is active, and $Z(m,k) = 0$ indicates that only noise is present in the given time-frequency slot (m,k).

The proposed method involves separate treatment of the instantaneous power of the observation, corresponding to the squared length of $\mathbf{X}$, and the spatial information contained in the orientation of the complex vector. Power and orientation vector are given by

$$A_1^2(m,k) := \|\mathbf{X}(m,k)\|^2 \tag{6}$$

$$\mathbf{Y}(m,k) := \mathbf{X}(m,k) / \|\mathbf{X}(m,k)\|, \tag{7}$$

where $\|\cdot\|$ denotes the Euclidean norm.

The first seed value is now chosen to be that observation, which most probably has the largest signal-to-noise power ratio (SNR). Without any prior knowledge about signal and noise other than the model of (4) and (5), this corresponds to finding the index m_1 of the observation with the largest power:

$$m_1(k) := \arg\max_m \left\{A_1^2(m,k)\right\}. \tag{8}$$

Exploiting the fact that for speech signals there is a strong temporal correlation in the activity patterns Z we define the local power spectral density (PSD) matrix by looking at the neighborhood of $\pm L$ frames around the frame index m_1:

$$\breve{\mathbf{\Phi}}_{\mathbf{YY},1}(k) := \frac{1}{2L+1} \sum_{m=m_1-L}^{m_1+L} \mathbf{Y}(m,k)\mathbf{Y}^H(m,k), \tag{9}$$

where $(\cdot)^H$ is the conjugate transpose operator. L should be chosen such that essentially only one source is dominant in this region.

As the seed value for the first column of the mixing matrix $\mathbf{H}$, i.e. the transfer function of the first source to the sensors, we suggest to take the eigenvector $\mathbf{v}_1$ corresponding to the largest eigenvalue of the local PSD matrix $\breve{\mathbf{\Phi}}_{\mathbf{YY},1}$:

$$\hat{\mathbf{H}}_1(k) := \mathbf{v}_1(k). \tag{10}$$

For the selection of the next seed vector $\hat{\mathbf{H}}_2$, we again want to choose an observation with large amplitude to ensure that the observation corresponds to speech. But in addition to that, the goal is to find an observation which corresponds to a different source than the previously chosen seed value. This second objective is realized by employing a weighting function on $A^2(m,k)$, which deemphasizes observations corresponding to sources of which seed values have already been drawn:

$$A_2^2(m,k) := A_1^2(m,k) \cdot \Psi_k\left(\hat{\mathbf{H}}_1(k), \mathbf{Y}(m,k)\right). \tag{11}$$

The frequency bin dependent weighting function Ψ_k should suppress observations with spatial properties which are similar to the already found seed vector $\hat{\mathbf{H}}_1$, whereas other observations should be unaffected. The optimal weighting function is therefore the probability that the observation $\mathbf{Y}(m,k)$ under consideration is from a different source than the already chosen one. The best estimate $\hat{Z}$ of the identity of the source which has produced the observation $\mathbf{Y}(m,k)$ is the value of Z that maximizes the posterior probability: $\hat{Z} = \arg\max_Z P(Z|Y)$. Let $z_1 := Z(m_1,k) \in \{1, \ldots, I\}$ denote the source that has produced $\mathbf{Y}(m_1,k)$. The weighting should then be according to

$$P\left(Z \neq z_1 | \mathbf{Y}\right) = 1 - P\left(Z = z_1 | \mathbf{Y}\right), \tag{12}$$

where here and in the following we leave out the arguments of $\mathbf{Y}$ and Z for ease of notation. Using Bayes' rule

$$P(Z = z_1|\mathbf{Y}) = \frac{p(\mathbf{Y}|Z = z_1)P(Z = z_1)}{\sum_{z'=1}^{I} p(\mathbf{Y}|Z = z')P(Z = z')} \tag{13}$$

and assuming all prior probabilities to be equal, we obtain

$$P(Z \neq z_1|\mathbf{Y}) = 1 - \frac{p\left(\mathbf{Y}|Z = z_1\right)}{\sum_{z'=1}^{I} p\left(\mathbf{Y}|Z = z'\right)}. \tag{14}$$

In previous work [1] we have proposed to model the conditional statistics of $\mathbf{Y}$ given Z by a circularly symmetric distribution defined on the complex hypersphere, namely the complex Watson probability density function (PDF) [8]. This PDF is given by

$$p\left(\mathbf{Y}|Z; \mathbf{F}_Z, \kappa_Z\right) := \frac{(D-1)!}{2\pi^D\, \underline{M}\left(1, D, \kappa_Z\right)}\; e^{\kappa_Z\, \left|\mathbf{F}_Z^H \mathbf{Y}\right|^2}, \tag{15}$$

where the parameter $\mathbf{F}_Z$ is the mean orientation, and the concentration parameter $\kappa_Z > 0$ characterizes the dispersion around the mean. $\underline{M}(a,b,z)$ is the confluent hypergeometric function of the first kind.

Of course, the parameters $\mathbf{F}_{z'}$, $\kappa_{z'}$ for $z' = 1, \ldots, I$ are not known since estimating these parameters is the goal of the BSS algorithm itself. Therefore, the following assumptions are made:

- $\mathbf{F}_{z_1} = \hat{\mathbf{H}}_1(k)$: the orientation of the first seed value is taken as an estimate of the orientation parameter of the density $p(\mathbf{Y}|Z = z_1)$ corresponding to source z_1.
- $\kappa_{z'} = \kappa$ for $z' = 1, \ldots, I$, i.e. the concentration parameters are assumed to be the same for all sources.
- The denominator of (14) is replaced by the likelihood at the mode of the PDF, i.e. $p\left(\mathbf{Y} = \mathbf{F}_{z_1}; \mathbf{F}_{z_1}, \kappa\right)$. This avoids the need to know the parameters of the other Watson distributions, while still achieving the required normalization.

With these assumptions the following weighting function is obtained:

$$\Psi\left(\hat{\mathbf{H}}_1, \mathbf{Y}\right) := 1 - \frac{p\left(\mathbf{Y}|Z = z_1; \hat{\mathbf{H}}_1, \kappa\right)}{p\left(\mathbf{Y} = \hat{\mathbf{H}}_1|Z = z_1; \hat{\mathbf{H}}_1, \kappa\right)}$$

$$= 1 - \exp\left(\kappa\left(\left|\hat{\mathbf{H}}_1^H \mathbf{Y}\right|^2 - 1\right)\right). \qquad (16)$$

Note that all parameters, and thus also the weighting function, are dependent on the frequency bin k although we neglected this dependence in our notation from eq. (12) on. The choice of the concentration parameter κ is not critical. A coarse estimate can be easily obtained by the maximum likelihood estimate of the concentration of all observations. Thus, we define the global PSD matrix by

$$\boldsymbol{\Phi}_{\mathbf{YY}}(k) := \frac{1}{M}\sum_{m=1}^{M} \mathbf{Y}(m,k)\mathbf{Y}^H(m,k). \qquad (17)$$

Following [8], the concentration parameter κ depends on the largest eigenvalue λ_1 of the PSD matrix $\boldsymbol{\Phi}_{\mathbf{YY}}$ and is computed by

$$\kappa := \eta_D^{-1}\left(\lambda_1\right), \qquad (18)$$

where the function $\eta_D\left(\cdot\right)$ is defined by

$$\eta_D(\kappa) := \frac{\underline{M}\left(2, D+1, \kappa\right)}{D \cdot \underline{M}\left(1, D, \kappa\right)}. \qquad (19)$$

The inverse function of (19) has to be computed by numerical methods [1].

The scheme described for the selection of the second seed value can now be repeated until I seed values have been found. For the $(i+1)$-th new seed the instantaneous powers $A_i^2(m,k)$ are weighted according to the previously chosen seed value $\hat{\mathbf{H}}_i$. The proposed seed selection algorithm is summarized in Alg. 1.

Algorithm 1 Proposed seeding algorithm

```
for all $k$ do
    for all $m$ do
        $A_1^2(m,k) := \|\mathbf{X}(m,k)\|^2$
        $\mathbf{Y}(m,k) := \mathbf{X}(m,k)\,/\,\|\mathbf{X}(m,k)\|$
    end for
    • Compute global PSD matrix $\boldsymbol{\Phi}_{\mathbf{YY}}(k)$ using (17).
    • Find largest eigenvalue $\lambda_1$ of $\boldsymbol{\Phi}_{\mathbf{YY}}(k)$.
    • Compute concentration parameter $\kappa$ using (18).
    for $i := 1$ to $I$ do
        • Find frame index $m_i$ of largest $A_i^2(m,k)$.
        • Compute local PSD matrix $\breve{\boldsymbol{\Phi}}_{\mathbf{YY},i}(k)$ using (9).
        • Compute principal eigenvector $\mathbf{v}_1(k)$ of $\breve{\boldsymbol{\Phi}}_{\mathbf{YY},i}(k)$.
        • Set seed value $\hat{\mathbf{H}}_i(k) := \mathbf{v}_1(k)$.
        for all $m$ do
            $A_{i+1}^2(m,k) := A_i^2(m,k) \cdot \Psi_k\left(\hat{\mathbf{H}}_i(k), \mathbf{Y}(m,k)\right)$
        end for
    end for
end for
```

4. Illustrative example

In the following we are going to illustrate the proposed seed value selection with a toy example with real-valued two-dimensional data ($D = 2$) and three sources ($I = 3$). Figure 1(a) shows a scatter plot of the observations and the sample with the largest amplitude chosen as the first seed value. For clarity of presentation we neglect the consideration of the neighborhood introduced in (9) and set $L = 0$. Then the principal eigenvector of the local PSD is equal to the observation itself, and (10) gives as the first seed value: $\hat{\mathbf{H}}_1(k) := \mathbf{Y}(m_1, k)$.

To illustrate the effect of the weighting function, we approximate the exponential in (16) by a Taylor series truncated after the linear term. Setting $\kappa = 1$ we obtain

$$\Psi'\left(\hat{\mathbf{H}}_1, \mathbf{Y}\right) := 1 - \left|\hat{\mathbf{H}}_1^H \mathbf{Y}\right|^2, \qquad (20)$$

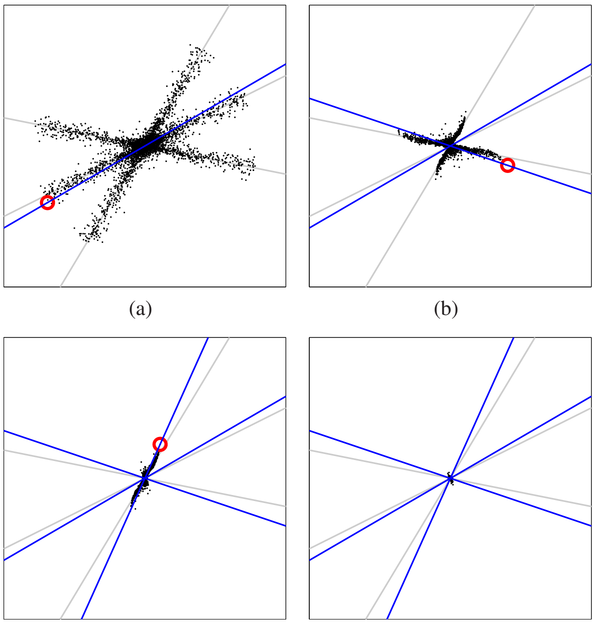
which is equal to the square of the sine of the angle between $\hat{\mathbf{H}}_1$ and $\mathbf{Y}$.

This weighting corresponds to a projection of the observation $\mathbf{X}(m,k)$ onto the orthogonal complement of the subspace spanned by $\hat{\mathbf{H}}_1$, as can be seen as follows:

$$\begin{aligned} A_2^2(m,k) &:= \left\|\left(\mathbf{I} - \hat{\mathbf{H}}_1(k)\hat{\mathbf{H}}_1^H(k)\right)\ \mathbf{X}(m,k)\right\|^2 \\ &= \left\|\left(\mathbf{I} - \hat{\mathbf{H}}_1(k)\hat{\mathbf{H}}_1^H(k)\right)\ \mathbf{Y}(m,k)A_1(m,k)\right\|^2 \\ &= A_1^2(m,k) \cdot \left(1 - \left|\hat{\mathbf{H}}_1^H(k)\mathbf{Y}(m,k)\right|^2\right). \qquad (21) \end{aligned}$$

Thus, if the candidate vector $\mathbf{Y}$ points to the same direction as the previous seed vector it receives a weight of zero and thus cannot be chosen as next seed value. This weighting is sensible, because an observation from a similar orientation as $\hat{\mathbf{H}}_1$ is likely to originate from the same source. With this deflation method directions of already chosen seed values are successively suppressed, as can be seen from Figs. 1(b)-(d).

Note that in contrast to the Gram-Schmidt procedure we just manipulate the vector lengths and do not remove the subspace spanned by the selected seed vector. In this way, we can obtain seed vectors which are not necessarily orthogonal to each other. This property allows us to apply the proposed seed selection method also to undetermined source separation, where the number sources I exceeds the numbers of microphones D.



(a)

(b)

(c)

(d)

Figure 1: *Real-valued toy data example: Scatter plot of the weighted observations $A_i \cdot \mathbf{Y}$ (black dots) for 3 iterations of the seed selection algorithm (a) – (d), $\mathbf{H}_i$ (light gray lines), selected seed sample (red circle), $\hat{\mathbf{H}}_i$ (blue lines).*

5. Simulation Results

We have evaluated the proposed algorithm on data taken from the development dataset with $D = 4$ and $I = 3$ of the "Source separation in the presence of real-world background noise" task of the second signal separation evaluation campaign (SiSEC2010) [9].

The performance is judged in terms of the similarity of the estimated transfer function vector $\hat{\mathbf{H}}_i$ to the true vector $\mathbf{H}_i$, where $\mathbf{H}_i$ is obtained as the principal eigenvector of the PSD matrix of the microphone signal if only the i-th source is present. We define the similarity as the squared inner product between these vectors, averaged over all sources and frequencies in the set $\mathcal{K}$ (corresponding to 70 Hz-7.2 kHz)

$$\mathcal{Q} := \frac{1}{I\,|\mathcal{K}|}\sum_{i=1}^{I}\sum_{k\in\mathcal{K}}\left|\hat{\mathbf{H}}_i^H(k)\mathbf{H}_i(k)\right|^2. \tag{22}$$

The similarity score $\mathcal{Q}$ varies between 0 and 1, where 1 is reached if the estimates perfectly match the orientation of the true transfer function vectors. Note that we solved the permutation problem before computation of (22) by applying a oracle permutation alignment method which utilized information of the sources.

Ideally, each source should be represented by exactly one seed vector. To measure how well this is fulfilled let

$$u_k(i) := \underset{l=1,\ldots,I}{\arg\max}\left|\hat{\mathbf{H}}_i^H(k)\mathbf{H}_l(k)\right|^2 \tag{23}$$

be the index of the orientation which is closest to the seed vector $\hat{\mathbf{H}}_i$. As the average allocation score we define

$$\mathcal{A} := \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}|\mathcal{J}_k| \text{ with } \mathcal{J}_k := \{u_k(i), i = 1,\ldots,I\}, \tag{24}$$

i.e., the average number of *different* indices in the sets $\mathcal{J}_k$. Clearly, an allocation score $\mathcal{A} = I$ corresponds to the case that all sources are represented by exactly one seed vector.

We compared the proposed seed selection to drawing samples randomly from the set of observation vectors (RS) and to the hierarchical agglomerative clustering (HAC) based seeding scheme proposed in [3]. While the computational complexity of RS is independent of the number of observations, the computational cost of the proposed method increases linearly and that of HAC quadratically with the number of observations. Hence, we have tested a power based subsampling of the observations before applying HAC to reduce computational cost and improve the quality of the seeds (HAC2).

The chosen seed values are used in our iterative EM-based BSS algorithm [1], and the above defined performance indicators are measured after each iteration of the EM algorithm. The averaged score over the 10 different setups of the SiSEC2010 task is depicted in Fig. 2.

As expected the performance of random sample seeding is the worst of all seeding schemes. Although HAC has the highest computational requirements its performance gains are moderate. This is caused by the count-based selection of the seeds at the end of the clustering. Therefore, the energy based subsampling prior to HAC achieved higher performance results.

We observed that the proposed algorithm consistently achieved the highest performance of the tested seeding schemes. The quality of the selected seeds is high and only few iterations are needed to arrive at the optimum. Furthermore, the final values of the performance measures are higher than with the other seed selection methods.

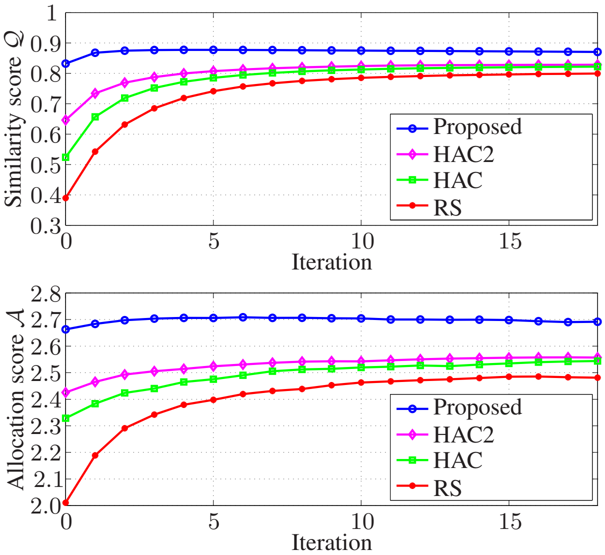


Figure 2: *Comparison of seed selection algorithms: Performance measures Eqs.* (22) *and* (24) *versus EM iteration count*

6. Conclusion

We introduced a blind seeding algorithm for iterative BSS with moderate computational requirements. We confirmed by simulations that the algorithm provides good seeds even for the challenging conditions of SiSEC2010 where a significant amount of noise is present. In future research we will explore an extension of the proposed algorithm to detect the number of sources in highly reverberant and noisy conditions.

7. References

[1] D. H. Tran-Vu and R. Haeb-Umbach, "An EM approach to integrated multichannel speech separation and noise suppression," in *Proc. IWAENC*, 2010.

[2] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, 2011.

[3] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, 2010.

[4] S. Araki, T. Nakatani, and H. Sawada, "Simultaneous clustering of mixing and spectral model parameters for blind sparse source separation," in *Proc. ICASSP*, 2010.

[5] Z. E. Chami, A. D.-T. Pham, C. Serviere, and A. Guerin, "A new EM algorithm for underdetermined convolutive blind source separation," in *Proc. EUSIPCO*, 2009.

[6] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," in *Proc. IEEE Int Acoustics, Speech, and Signal Processing, 1993. ICASSP-93. Conf*, 2002, vol. 1.

[7] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ica and beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, 2006.

[8] K. V. Mardia and I. L. Dryden, "The complex watson distribution and shape analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 4, 1999.

[9] F. Theis, G. Nolte, and S. Araki, "Signal separation evaluation campaign," *http://sisec.wiki.irisa.fr/*.