Estimating the Number of Sources for Frequency-Domain Blind Source Separation

Hiroshi Sawada, Stefan Winter*, Ryo Mukai, Shoko Araki, and Shoji Makino

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

Abstract. Blind source separation (BSS) for convolutive mixtures can be performed efficiently in the frequency domain, where independent component analysis (ICA) is applied separately in each frequency bin. To solve the permutation problem of frequency-domain BSS robustly, information regarding the number of sources is very important. This paper presents a method for estimating the number of sources from convolutive mixtures of sources. The new method estimates the power of each source or noise component by using ICA and a scaling technique to distinguish sources and noises. Also, a reverberant component can be identified by calculating the correlation of component envelopes. Experimental results for up to three sources show that the proposed method worked well in a reverberant condition whose reverberation time was 200 ms.

1 Introduction

Blind source separation (BSS) [1] is a technique for estimating original source signals solely from their mixtures at sensors. In some applications, such as audio acoustics, signals are mixed in a convolutive manner with reverberations. This makes the BSS problem more difficult to solve than an instantaneous mixture problem. Let us formulate the convolutive BSS problem. Suppose that N source signals $s_k(t)$ are mixed and observed at M sensors

$$x_j(t) = \sum_{k=1}^{N} \sum_l h_{jk}(l) s_k(t-l) + n_j(t), \tag{1}$$

where $h_{jk}(l)$ represents the impulse response from source k to sensor j and $n_j(t)$ is an additive Gaussian noise for each sensor. The goal is to obtain N output signals $y_i(t)$, each of which is a filtered version of a source $s_k(t)$. If we have enough sensors ($M \geq N$), a set of FIR filters $w_{ij}(l)$ of length L is typically used to produce separated signals

$$y_i(t) = \sum_{j=1}^{M} \sum_{l=0}^{L-1} w_{ij}(l) x_j(t-l) \tag{2}$$

at the outputs, and independent component analysis (ICA) [2] is generally used to obtain the FIR filters $w_{ij}(l)$. If the number of sensors is insufficient ($M < N$),

* The author is on leave from the Chair of Multimedia Communications and Signal Processing, University Erlangen-Nuremberg

C.G. Puntonet and A. Prieto (Eds.): ICA 2004, LNCS 3195, pp. 610–617, 2004.

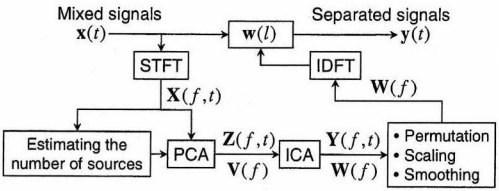

Fig. 1. Flow of frequency-domain BSS.

we need to rely on the sparseness of source signals, and approaches for separation become totally different [3]. Therefore, this paper focuses on cases where we have enough sensors ($M \geq N$).

There are two major approaches to the convolutive BSS problem. The first is time-domain BSS, where ICA is applied directly to the convolutive mixture model [4, 5]. The other approach is frequency-domain BSS, where complex-valued ICA for instantaneous mixtures is applied separately in each frequency bin [6–10]. The merit of frequency-domain BSS is that ICA for instantaneous mixtures is simpler and computationally more efficient than ICA for convolutive mixtures. We have implemented a frequency-domain BSS system that can separate three sources in real-time [10]. The price we must pay for this computational efficiency includes several additional problems that need to be solved for integrating the ICA solutions obtained separately in each frequency bin. The permutation problem is the best known. The permutation ambiguity of ICA should be aligned so that a separated signal in the time-domain contains the frequency components of the same source signal. We have proposed a method for solving the permutation problem [8], which performs well even if the number of sources is large [9, 10]. However, this method requires knowledge of the number of sources, and we assumed that the number was known apriori in these papers.

In this paper, we propose a method for estimating the number of sources N in the context of frequency-domain BSS. It is well known that the number of dominant eigenvalues of the spatial correlation matrix corresponds to the number of sources [11, 12]. However, it is difficult to decide whether an eigenvalue is dominant or not for sensor observations mixed in a reverberant condition as shown in Sec. 3. This difficulty has already been pointed out in [12], where they propose the use of support vector machines (SVM) to classify eigenvalue distributions and decide the number of sources. However, the SVM needs to be trained beforehand and experimental results were provided only for 1- or 2-source cases. Our proposed method is based on an analysis of ICA solutions obtained in the frequency domain as shown in Sec. 4. Experimental results for up to three sources show that the method worked well in a real reverberant condition.

2 Frequency-Domain BSS

This section describes frequency-domain BSS whose flow is shown in Fig. 1. First, time-domain signals $x_j(t)$ at sensors are converted into frequency-domain

time-series signals $X_j(f,t)$ by short-time Fourier transform (STFT), where t is now down-sampled with the distance of the frame shift. Then, the number of sources N should be estimated from $\mathbf{X}(f,t) = [X_1(f,t), \ldots, X_M(f,t)]^T$. This part is the main topic of this paper, and will be discussed in Secs. 3 and 4.

After estimating the number of sources N, the dimension M of sensor observations $\mathbf{X}(f,t)$ is reduced to N typically by principal component analysis (PCA), $\mathbf{Z}(f,t) = \mathbf{V}(f)\mathbf{X}(f,t)$, where $\mathbf{V}(f)$ is an $N \times M$ matrix whose row vectors generate N principal components [13]. Even if $N = M$, PCA is useful as preprocessing. Then, complex-valued ICA $\mathbf{Y}(f,t) = \mathbf{B}(f)\mathbf{Z}(f,t)$ is applied, where $\mathbf{B}(f)$ is an N dimensional square matrix. Through these operations, the sensor observations $\mathbf{X}(f,t)$ are separated into independent components $\mathbf{Y}(f,t) = [Y_1(f,t), \ldots, Y_N(f,t)]^T$ by $\mathbf{Y}(f,t) = \mathbf{W}(f)\mathbf{X}(f,t)$, where $\mathbf{W}(f) = \mathbf{B}(f)\mathbf{V}(f)$. Note that $\mathbf{W}(f)$ is invertible if $\mathbf{V}(f)$ is full rank and $\mathbf{B}(f)$ is made unitary (by e.g. FastICA [2]).

The ICA solution $\mathbf{W}(f)$ in each frequency bin has permutation and scaling ambiguity: even if we permute the rows of $\mathbf{W}(f)$ or multiply a row by a constant, it is still an ICA solution. In matrix notation, $\mathbf{\Lambda}(f)\mathbf{P}(f)\mathbf{W}(f)$ is also an ICA solution for any permutation $\mathbf{P}(f)$ and diagonal $\mathbf{\Lambda}(f)$ matrix. The permutation ambiguity $\mathbf{P}(f)$ should be solved so that $Y_i(f,t)$ at all frequencies corresponds to the same source $s_i(t)$. We use the method described in [8]. The scaling ambiguity $\mathbf{\Lambda}(f)$ can be solved by making $Y_i(f,t)$ as close to a part of the sensor observation $\mathbf{X}(f,t)$ as possible. The minimal distortion principle (MDP) [4] makes $y_i(t)$ as close to $\sum_l h_{ii}(l)s_i(t-l)$, a part of $x_i(t)$, as possible. In the frequency domain, it is realized by $\mathbf{\Lambda}(f) = \mathrm{diag}[\mathbf{W}^{-1}(f)]$ [7]. If $N < M$, the Moore-Penrose pseudoinverse $\mathbf{W}^{+}(f)$ is used instead of $\mathbf{W}^{-1}(f)$. Also, the scaling (3) that will be discussed in Sec. 4 can be used.

The aligned matrices $\mathbf{W}(f) \leftarrow \mathbf{\Lambda}(f)\mathbf{P}(f)\mathbf{W}(f)$ are the frequency responses of separation filters $\mathbf{w}(l)$. However, we need to be concerned about the circularity effect of discrete frequency representation. We perform spectral smoothing [14] for $[\mathbf{W}(f)]_{ij}$ to mitigate the circularity effect. Finally, time-domain filters $w_{ij}(l)$ are obtained by applying inverse DFT to the smoothed elements $[\mathbf{W}(f)]_{ij}$.

3 Conventional Eigenvalue-Based Method

This section describes a conventional eigenvalue-based method for estimating the number of sources in each frequency bin [11]. It performs eigenvalue decomposition for the spatial correlation matrix $\mathbf{R}(f) = \langle \mathbf{X}(f,t)\mathbf{X}(f,t)^H \rangle_t$ of sensor observations, where $\langle \cdot \rangle_t$ denotes the averaging operator and $\cdot^H$ denotes the conjugate transpose. Let $\lambda_1 \geq \cdots \geq \lambda_N \geq \cdots \geq \lambda_M$ be the sorted eigenvalues of $\mathbf{R}(f)$. If there is no reverberation, the number of dominant eigenvalues is equal to the number of sources N, and the remaining $M - N$ smallest eigenvalues are the same as the noise power: $\lambda_{N+1} = \cdots = \lambda_M = \sigma_n^2$. However, there are two problems in a real reverberant condition.

Reverberation. The number of dominant eigenvalues might be more than the number of source signals, if the reverberation of a mixing system is long and

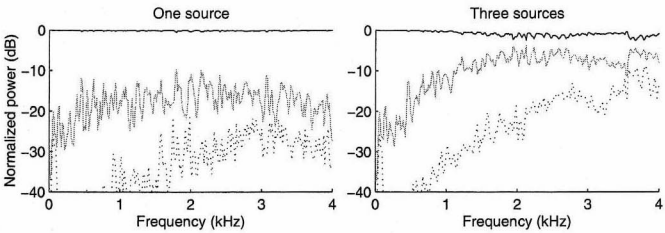

Fig. 2. Component powers estimated by the eigenvalue-based method.

strong. This is because the reverberation of a mixing system, i.e. the non-zero part of $h_{jk}(l)$, is usually longer than the STFT frame, and the reverberation component could be counted as a signal.

Unrecovered power. The number of dominant eigenvalues might be less than the number of source signals, if some of the column vectors of the mixing matrix are similar. In this case, the first few eigenvalues represent almost all powers. A typical situation can be seen in low frequencies, where the phase differences among sensors are very small.

Because of these two problems, the eigenvalue-based method does not work well in a real reverberant condition. Figure 2 shows component powers estimated by the eigenvalue-based method in an environment whose conditions are summarized in Fig. 3. The left hand plot shows a one-source case. Because of reverberations, the normalized power of the second principal components were around -20 dB. To distinguish the source and noises (including reverberations), a threshold of around -15 dB is good for the one-source case. However, if such a threshold is used for the three-source case shown in the right hand plot, the number of sources is estimated at two in most frequency bins. Therefore, it is hard to find a threshold that works well for both cases.

4 Proposed ICA-Based Method

In this section, we propose a new method for estimating the number of sources that solves the two problems mentioned above.

To solve the problem of unrecovered power, the proposed method recovers the power of each signal measured at sensors by using ICA and a scaling technique. It first applies ICA for $\mathbf{X}(f,t)$ without performing dimension reduction, i.e. assuming the number of sources N is equal to the number of sensors M. Because of the scaling ambiguity of ICA, the power of each component of the ICA solution $\mathbf{Y}(f,t) = \mathbf{W}(f)\mathbf{X}(f,t)$ is different from the power of each source or noise. If the real number of sources is less than M, $M-N$ noise components are generally enhanced.

To recover the power of each component measured at sensors, we use a scaling

$$\mathbf{\Lambda}(f) = \mathrm{sqrt}(\mathrm{diag}[\mathbf{W}^{-H}(f)\mathbf{W}^{-1}(f)]), \tag{3}$$

for ICA solution $\mathbf{Y}(f,t) = \mathbf{\Lambda}(f)\mathbf{W}(f)\mathbf{X}(f,t)$. Note again that $\mathbf{W}(f)$ is invertible if the smallest eigenvalue of the spatial correlation matrix $\mathbf{R}(f)$ is not zero. We call (3) *power-recovery scaling* since it recovers the power of the sensor observations as follows. Firstly, the total power of sensor observations is recovered at the outputs:

$$||\mathbf{Y}(f,t)||^2 = ||\mathbf{X}(f,t)||^2, \tag{4}$$

if the components of $\mathbf{Y}(f,t)$ are uncorrelated. Moreover, if ICA is properly solved and Y_i's are made mutually independent, the power of each source measured at sensors is recovered at each output:

$$|Y_i(f,t)|^2 = ||\mathbf{H}_{\Pi(i)}(f)S_{\Pi(i)}(f,t)||^2, \tag{5}$$

where Π is a permutation, S_k is the k-th source and $\mathbf{H}_k$ is the mixing vector of S_k. This equation (5) can be seen as decomposition of equation (4). We have proved both equations. However, the proofs are omitted here for space limit.

In this way, the power of each component $Y_i(f,t)$ of the ICA solution $\mathbf{Y}(f,t) = \mathbf{\Lambda}(f)\mathbf{W}(f)\mathbf{X}(f,t)$ approaches the real power of each source or noise measured at sensors $\mathbf{H}_{\Pi(i)}(f)S_{\Pi(i)}(f,t)$. Therefore, the power

$$\sigma_i^2 = \langle |Y_i(f,t)|^2 \rangle_t \tag{6}$$

can be used as a criterion for distinguishing sources and noises (including reverberations). Although the MDP explained in Sec. 2 can also be used for power estimation, the power recovered by the MDP contains only the power of a selected sensor $x_i(t)$, and is sensitive to the sensor selection. The power recovered by the power-recovery scaling (3) contains the power of all sensors, and is therefore more robust for power estimation.

The problem of reverberation discussed in Sec. 3 still needs to be solved. We observed that the envelope of a reverberant component has a strong correlation with the envelope of a source component. The correlation of two envelopes $|Y_{i1}(f,t)|$ and $|Y_{i2}(f,t)|$, $i1, i2 \in \{1,\ldots,M\}$, is defined as

$$\frac{\langle v_{i1}(t) \cdot v_{i2}(t) \rangle_t}{\sqrt{\langle v_{i1}^2(t) \rangle_t} \cdot \sqrt{\langle v_{i2}^2(t) \rangle_t}}, \quad \text{where } v_i(t) = |Y_i(f,t)| - \langle |Y_i(f,t)| \rangle_t. \tag{7}$$

When $Y_{i1}(f,t)$ is a source component and $Y_{i2}(f,t)$ is not a source component but includes the reverberation of source $i1$, the correlation of $|Y_{i1}(f,t-\Delta t)|$ and $|Y_{i2}(f,t)|$ with an appropriate time delay $-\Delta t$ tends to be large. Therefore, the correlation can be used as a measure to distinguish sources and reverberations.

The overall procedure of the proposed method is as follows.

1. Calculate independent components $Y_i(f,t)$ by using ICA and scaling (3).
2. If the normalized power $\sigma_i^2 / \sum_{k=1}^{M} \sigma_k^2$ of i-th component is smaller than a threshold, e.g. 0.01 (-20 dB), consider it to be a noise component.
3. If the normalized power $\sigma_i^2 / \sum_{k=1}^{M} \sigma_k^2$ is smaller than a threshold, e.g. 0.2, and one of the correlations (7) among other components is larger than a threshold, e.g. 0.5, consider it to be a reverberant component.
4. Otherwise, consider the i-th component to be a signal.

These thresholds can be determined beforehand by the power levels of background noise and reverberations.

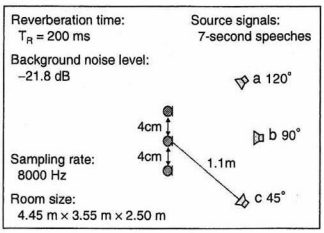

Fig. 3. Experimental conditions.

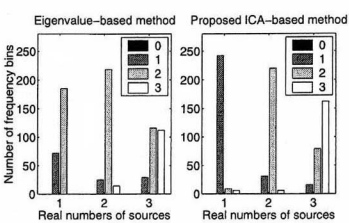

Fig. 4. Estimated numbers of sources.

Table 1. BSS results obtained with different estimation methods: the conventional eigenvalue-based method (Eig.) and the proposed ICA-based method (Prop.).

#sources (real)	1 (c)			2 (a,c)		3 (a,b,c)			
#sources (est.)	2 (Eig.)		1 (Prop.)	2 (Both)		2 (Eig.)		3 (Prop.)		
SIR (dB)	∞	∞	∞	17.1	17.1	2.0	−1.2	13.6	15.4	13.3
SDR (dB)	10.1	−4.4	∞	13.3	14.2	0.7	−3.5	9.4	10.5	10.2

5 Experimental Results

We performed experiments to estimate the number of sources from sensor observations and to separate them into source signals. Sensor observations were generated by convolving source signals with impulse responses and then adding background noise. The impulse responses and the background noise were measured in the conditions summarized in Fig. 3. We tested cases of one, two and three sources, while the number of sensors was three for all cases. Figure 4 shows the numbers of sources estimated by using the conventional eigenvalue-based method and the proposed ICA-based method. The vertical axis shows the number of frequency bins for each estimated number of sources. The STFT frame size was 512, and thus the number of total frequency bins to cover 0–4000 Hz was 257. By taking the maximum vote, the ICA-based method successfully estimated the number of sources in all cases, whereas the eigenvalue-based method estimated the number of sources at 2 in all cases.

Table 1 shows the BSS results obtained with these estimations for the number of sources. The results were measured in terms of the signal-to-interference ratio (SIR) and signal-to-distortion ratio (SDR) of each output. To calculate the SIR of $y_i(t)$, it is decomposed as $y_i(t) = tar_i(t) + int_i(t)$, where $tar_i(t)$ is a filtered component of a target signal $s_{\Pi(i)}(t)$ and $int_i(t)$ is the remaining interference component. The SIR is the power ratio of $tar_i(t)$ and $int_i(t)$. The mapping Π was selected to maximize the SIR. To calculate the SDR of $y_i(t)$, the filtered component of the target signal is further decomposed as $tar_i(t) = \alpha_i \cdot ref_i(t) + e_i(t)$, where $ref_i(t)$ is a reference signal and α_i is a scalar that minimizes the error power of $e_i(t)$. We used $ref_i(t) = \sum_l h_{ii}(l)s_i(t-l)$ following the MDP. The SDR is the power ratio of $\alpha_i \cdot ref_i(t)$ and $e_i(t)$. The BSS performance was degraded if

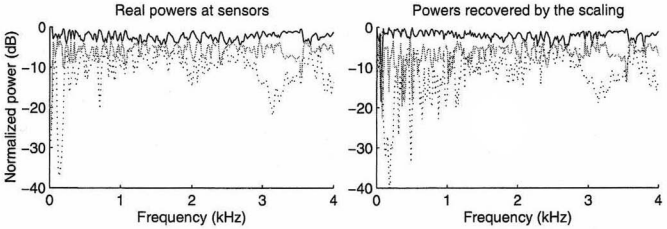

Fig. 5. Power recovery by scaling formula (3) when there are three sources.

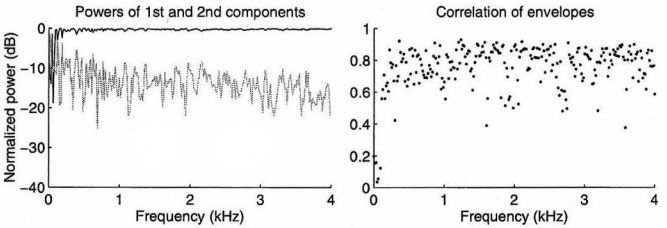

Fig. 6. Identifying reverberant components when there is one source.

the number of sources was incorrectly estimated. In the one-source case with the eigenvalue-based method, the number of sources was overestimated. Thus the source was decomposed into two outputs, and the SDRs were poor. In the three-source case, again with the eigenvalue-based method, the number of sources was underestimated. In this case, the output signals were still mixed, and thus the SIRs as well as SDRs were poor.

Figure 5 shows how well the powers of sources were recovered by ICA and the proposed scaling technique. The left hand plot shows the normalized powers of the three sources measured at sensors, and the right hand plot shows those estimated by ICA and the scaling formula (3). The powers were sufficiently well recovered to estimate the number of sources. Compared with the result obtained with the eigenvalue-based method (the right hand plot in Fig. 2), the advantage of the proposed method becomes clear.

Figure 6 shows how the proposed method copes with the reverberation problem. This case had only one source. The left hand plot shows the normalized power of the first and second largest components of the scaled ICA outputs. It was hard to decide solely from these normalized powers whether the second component was a signal or a noise because the powers of the second components were not sufficiently small in many frequency bins. However, by calculating the correlation of the envelopes between the first and second components, it became clear that the second component was a reverberation, i.e. a noise. The right hand plot shows the correlations, which were large enough (around 0.8) in many frequency bins.

6 Conclusion

We have proposed a method for estimating the number of sources in each frequency bin. Our method provides a solution for the two problems with the conventional eigenvalue-based method discussed in Sec. 3, and provides a good estimation even in a reverberant condition of $T_R = 200$ ms. With the proposed method, frequency-domain BSS can be practically applied without apriori knowledge of the number of sources.

References

1. Haykin, S., ed.: Unsupervised adaptive filtering (Volume I: Blind source separation). John Wiley & Sons (2000)
2. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. John Wiley & Sons (2001)
3. Rickard, S., Balan, R., Rosca, J.: Blind source separation based on space-time-frequency diversity. In: Proc. ICA2003. (2003) 493–498
4. Matsuoka, K., Nakashima, S.: Minimal distortion principle for blind source separation. In: Proc. ICA 2001. (2001) 722–727
5. Douglas, S.C., Sun, X.: Convolutive blind separation of speech mixtures using the natural gradient. Speech Communication **39** (2003) 65–78
6. Smaragdis, P.: Blind separation of convolved mixtures in the frequency domain. Neurocomputing **22** (1998) 21–34
7. Murata, N., Ikeda, S., Ziehe, A.: An approach to blind source separation based on temporal structure of speech signals. Neurocomputing **41** (2001) 1–24
8. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech and Audio Processing **12** (2004)
9. Mukai, R., Sawada, H., de la Kethulle, S., Araki, S., Makino, S.: Array geometry arrangement for frequency domain blind source separation. In: Proc. IWAENC2003. (2003) 219–222
10. Sawada, H., Mukai, R., Araki, S., Makino, S.: Convolutive blind source separation for more than two sources in the frequency domain. In: Proc. ICASSP 2004. (2004)
11. Wax, M., Kailath, T.: Detection of signals by information theoretic criteria. IEEE Trans. Acoustics, Speech, and Signal Processing **33** (1985) 387–392
12. Yamamoto, K., Asano, F., van Rooijen, W., Ling, E., Yamada, T., Kitawaki, N.: Estimation of the number of sound sources using support vector machines and its application to sound source separation. In: Proc. ICASSP 2003. (2003) 485–488
13. Winter, S., Sawada, H., Makino, S.: Geometrical interpretation of the PCA subspace method for overdetermined blind source separation. In: Proc. ICA2003. (2003) 775–780
14. Sawada, H., Mukai, R., de la Kethulle, S., Araki, S., Makino, S.: Spectral smoothing for frequency-domain blind source separation. In: Proc. IWAENC2003. (2003) 311–314