Blind Separation of More Speech than Sensors with Less Distortion by Combining Sparseness and ICA

Shoko Araki *Shoji Makino* *Audrey Blin* *Ryo Mukai* *Hiroshi Sawada*

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: shoko@cslab.kecl.ntt.co.jp

Abstract

We propose a method for separating speech signals with little distortion when the signals outnumber the sensors. Several methods have already been proposed for solving the underdetermined problem, and some of these utilize the sparseness of speech signals. These methods employ binary masks that extract a signal at time points where the number of active sources is estimated to be only one. However, these methods result in an unexpected excess of zero-padding and so the extracted speeches are severely distorted and have loud musical noise. In this paper, we propose combining a sparseness approach and independent component analysis (ICA). First, using sparseness, we estimate the time points when only one source is active. Then, we remove this single source from the observations and apply ICA to the remaining mixtures. Experimental results show that our proposed sparseness and ICA (SPICA) method can separate signals with little distortion even in a reverberant condition.

1. Introduction

Blind source separation (BSS) is an approach that estimates original source signals $s_i(n)$ using only information on the mixed signals $x_j(n)$ observed in each input channel. This technique can be used for noise robust speech recognition and high-quality hearing aid systems. It may also become a clue to auditory scene analysis.

In this paper, we consider the BSS of speech signals observed in a real environment, i.e., the BSS of convolutive mixtures of speech. We focus particularly on the underdetermined BSS problem, that is, the case of the number of source signals outnumbers the number of sensors.

Several methods have already been proposed for solving the underdetermined problem and some of these utilize the sparseness of speech signals [1, 2, 3]. If the signals are sparse enough, that is, most of the samples of a signal are almost zero, we can assume that the sources rarely overlap. Sparseness approaches use this assumption and extract each signal using time-frequency binary masks. However, due to these binary masks, their methods result in too much zero-padding to the extracted signals, and so the extracted speeches are severely distorted and sound unnatural.

To overcome this problem, i.e., to reduce the distortion of the extracted signals, we propose combining a sparseness approach and independent component analysis (ICA). First, using sparseness, we estimate the time points when only one source is active. Then, instead of extracting only one signal from the observations, we remove this single source from the observations and apply ICA to the remaining mixtures in order to separate the signals. This removal does not cause severe zero-padding to the separated signals, therefore we can improve the sound quality of the separated signals. Experimental results show that our sparseness and ICA (SPICA) method can separate signals with little distortion even in an echoic environment.

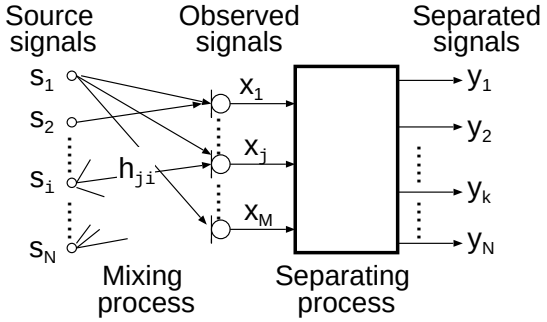

Figure 1: Block diagram of underdetermined BSS. $N > M$.

2. Problem description

In real environments, N signals observed by M sensors are modeled as convolutive mixtures $x_j(n) = \sum_{i=1}^{N} \sum_{k=1}^{P} h_{ji}(k) s_i(n-k+1)$ $(j = 1, \cdots, M)$, where s_i is the signal from a source i, x_j is the signal observed by a sensor j, and h_{ji} is the P-taps impulse response from a source i to a sensor j (see Fig. 1). Here, we consider the underdetermined case $N > M$. In this paper $N = 3$ and $M = 2$. Sources are assumed to be mutually independent and sparse.

This paper employs a time-frequency domain approach because speech signals are more sparse in the time-frequency domain [3, 4] and we can convert convolutive mixture problems into instantaneous mixture problems in each frequency. In the time-frequency domain, mixtures are modeled as $\boldsymbol{X}(\omega, m) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega, m)$, where $\boldsymbol{H}(\omega)$ is a 2×3 mixing matrix whose j-i component is a transfer function from a source i to a sensor j and $\boldsymbol{S}(\omega, m) = [S_1(\omega, m), S_2(\omega, m), S_3(\omega, m)]^T$, $\boldsymbol{X}(\omega, m) = [X_1(\omega, m), X_2(\omega, m)]^T$ and $\boldsymbol{Y}(\omega, m) = [Y_1(\omega, m), Y_2(\omega, m), Y_3(\omega, m)]^T$ show a Fourier transformed

source, observed and separated signals, respectively. ω is the frequency and m is the frame index.

Our objective is to estimate separated signals $\boldsymbol{Y}(\omega, m)$ using only the information provided by observations $\boldsymbol{X}(\omega, m)$. In this paper, the sources are speech signals, i.e., the sources are sufficiently sparse in the time-frequency domain [1]-[4].

3. Conventional methods with sparseness

The standard ICA cannot be applied in underdetermined cases because it assumes that a mixing matrix is invertible. Several separation methods employing source sparseness have been proposed for use when there are more sources than sensors [1, 2, 3].

If most of the samples of a signal are almost zero, we say that this signal is sparse. When signals are sparse enough, we can assume that the sources overlap at rare intervals. We can assume the sparseness of the speech signals especially in the time-frequency domain. For a detailed analysis of sparseness, see [5].

Sparseness approaches use this assumption and extract each signal using time-frequency binary masks. Because we can assume that sources do not overlap very often, we can extract each source by selecting the time points at which there is only one signal. One way of estimating such time points is to use the level difference of the observations and the phase difference between the observations. In this paper, we utilize omni-directional microphones, and we use the phase difference $\varphi(\omega, m) = \angle \frac{X_1(\omega,m)}{X_2(\omega,m)}$ between the observations $X_1(\omega, m)$ and $X_2(\omega, m)$.

Using $\varphi(\omega, m)$, we can estimate the direction of arrival (DOA) for each time point m by calculating $\theta(\omega, m) = \cos^{-1} \frac{\varphi(\omega,m)c}{\omega d}$, where c is the speed of sound and d is the microphone spacing. When we plot this DOA $\theta(\omega, m)$, we can see three peaks in the histogram for each frequency. Let these peaks be $\tilde{\theta}_1, \tilde{\theta}_2$ and $\tilde{\theta}_3$ where $\tilde{\theta}_1 \leq \tilde{\theta}_2 \leq \tilde{\theta}_3$, and the signal from $\tilde{\theta}_\xi$ be $\tilde{S}_\xi (\xi = 1, 2, 3)$ (Fig. 2).

We can extract each signal with a binary mask

$$M_\xi(\omega, m) = \begin{cases} 1 & \tilde{\theta}_\xi - \Delta \leq \theta(\omega, m) \leq \tilde{\theta}_\xi + \Delta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

by calculating $Y_\xi(\omega, m) = M_\xi(\omega, m) X_j(\omega, m)$ where j=1 or 2. Here, Δ is an extraction range parameter: if Δ is small the separation performance is good but the distortion (musical noise) becomes large, on the other hand, if Δ is large the musical noise problem is overcome but the separation performance deteriorates.

Although we can extract each signal using this binary mask (1), such extracted signals are discontinuously zero-padded by the binary mask. Therefore, we can hear considerable musical noise in the extracted output.

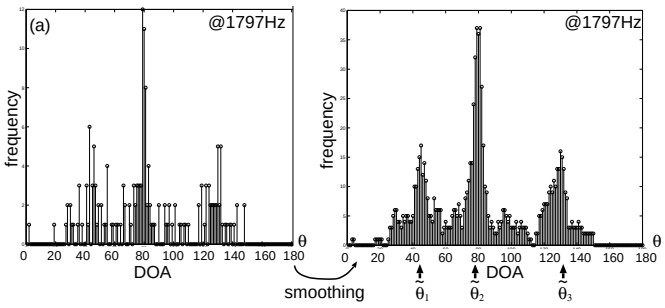

Figure 2: Example of (a) histogram and (b) smoothed histogram. Anechoic, female-male-male combination. DFTsize T =512.

4. Proposed Method: Combination of sparseness and ICA (SPICA)

To overcome this musical noise problem, we propose using both sparseness and ICA. Our method has two stages (see Fig. 3). In the first stage, contrary to the conventional approach, we *remove* one source from mixtures using the signals' sparseness. By this removal, it is expected that their zero-padding to be unimportant because we extract more time-frequency points than the conventional approach. Moreover we can expect the remaining mixtures to consist of only two signals. Therefore, in the second stage, we can apply ICA (e.g., [6]) to these remaining mixtures. Because these separated signals are not highly zero-padded, we can expect less musical noise.

[1st stage] One source removal:

Here, we utilize omni-directional microphones, therefore, we use the phase difference between the observations to set $\tilde{\theta}_1, \tilde{\theta}_2$ and $\tilde{\theta}_3$ shown in the previous section.

Instead of extracting each source as in conventional approaches, we remove only one source from the mixtures with a binary mask

$$M^{pq}_{\text{ICA}}(\omega, m) = \begin{cases} 1 & \theta_{min} \leq \theta(\omega, m) \leq \theta_{max} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

by calculating

$$\hat{\boldsymbol{X}}(\omega, m) = M^{pq}_{\text{ICA}}(\omega, m) \boldsymbol{X}(\omega, m). \quad (3)$$

In (2), θ_{min} and θ_{max} are extraction range parameters, and in (3), $\hat{\boldsymbol{X}}(\omega, m)$ is expected to be mixtures of $\tilde{S}_p$ and $\tilde{S}_q$. For instance, as in case 1, if $\tilde{S}_1$ can be removed from the observations with a mask M^{23}_{ICA} we can use ICA to separate $\tilde{S}_2$ and $\tilde{S}_3$ in the next stage (see Fig. 3). In this case θ_{min} and θ_{max} in (2) can be $\tilde{\theta}_1 < \theta_{th1} = \theta_{min} < \tilde{\theta}_2, \theta_{max} = 180^\circ$ (see Fig. 4). Similarly in case 2, when $\tilde{S}_3$ is to be removed from the observations with a mask M^{12}_{ICA}, $\theta_{min} = 0^\circ, \tilde{\theta}_2 < \theta_{th2} = \theta_{max} < \tilde{\theta}_3$.

Because our system has only two outputs, both removals should be performed to obtain three separated signals (see Fig. 3).

[2nd stage] Separation of remaining sources by ICA:

Because the remaining signals $\hat{\boldsymbol{X}}$ are expected to be mixtures of two signals, we can use 2×2 ICA to separate

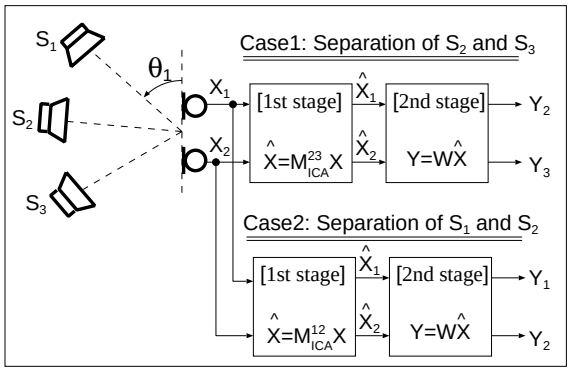

Figure 3: System setup

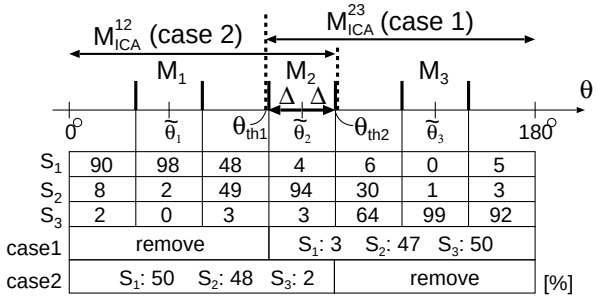

Figure 4: Source power in each area (%).

$\hat{\boldsymbol{X}}$. The separation process can be formulated as $\boldsymbol{Y}(\omega, m) = \boldsymbol{W}(\omega)\hat{\boldsymbol{X}}(\omega, m)$, where $\hat{\boldsymbol{X}}$ is the masked observed signal obtained by (3), $\boldsymbol{Y}(\omega, m) = [Y_{\xi_1}(\omega, m), Y_{\xi_2}(\omega, m)]^T$ (ξ_1, ξ_2 =1,2,3) is the separated output signal, and $\boldsymbol{W}(\omega)$ represents a (2×2) unmixing matrix. $\boldsymbol{W}(\omega)$ is determined so that $Y_{\xi_1}(\omega, m)$ and $Y_{\xi_2}(\omega, m)$ become mutually independent. This calculation is carried out independently at each frequency.

In this paper, the adaptive rule is $\boldsymbol{W}_{i+1}(\omega) = \boldsymbol{W}_i(\omega) + \eta\left[\mathrm{diag}\left(\langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right) - \langle\Phi(\boldsymbol{Y})\boldsymbol{Y}^{\mathrm{H}}\rangle\right]\boldsymbol{W}_i(\omega)$, where $\Phi(\boldsymbol{y}) = \phi(|\boldsymbol{y}|)\cdot e^{j\cdot\angle(\boldsymbol{y})}$, $\phi(x) = \tanh(gx)$ and $g = 100$ [7]. For solving the permutation problem of frequency domain ICA, we utilized the direction of arrival approach [8], and for solving scaling problem of frequency domain ICA, we used the minimum distortion principle [9].

5. Experiments

5.1. Experimental conditions

For the anechoic tests, we simulated the recording in an anechoic room using the mixing matrix $H_{ji}(\omega) = \exp(j\omega\tau_{ji})$, where $\tau_{ji} = \frac{d_j}{c}\sin\theta_i$, d_j is the position of the j-th microphone, θ_i is the direction of the i-th source, and c is the speed of sound. We simulated a pair of omni-directional microphones with an inter-element spacing of 4 cm. The sampling rate was 8 kHz. The speech signals arrived from three directions, $50°$(female), $100°$(male) and $135°$(male).

For the reverberant tests, we recorded each speech signal in a real room whose reverberation time was T_R=130 ms

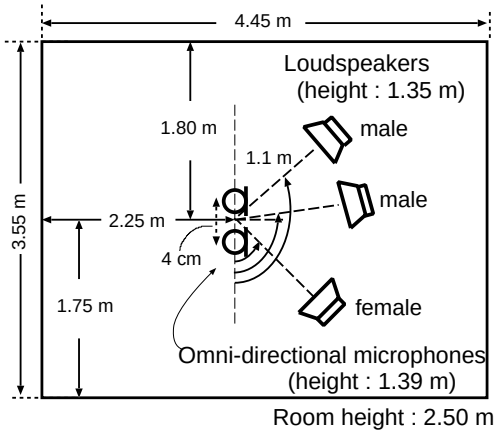

Figure 5: Room for echoic tests.

(Fig. 5) and added them to obtain the mixtures.

We constructed a histogram of $\theta(\omega, m)$ for each frequency, and smoothed it slightly (see Fig. 2). Then we divided the histogram into three parts, determined the peak for each part, and set these peaks as $\tilde{\theta_1}$, $\tilde{\theta_2}$ and $\tilde{\theta_3}$. The DFT frame size T was 512 and we used a frame shift of 256.

5.2. Performance measures

To evaluate the effectiveness of our approach, we used the signal to interference ratio (SIR) as a measure of separation performance, and the signal to distortion ratio (SDR) as a measure of sound quality:

$$\mathrm{SIR}_i = 10\log\frac{\sum_n y_{is_i}^2(n)}{\sum_{i\neq j}\sum_n y_{is_j}^2(n)} \quad (4)$$

$$\mathrm{SDR}_i = 10\log\frac{\sum_n x_{ks_i}^2(n)}{\sum_n (x_{ks_i}(n) - \alpha y_{is_i}(n-D))^2} \quad (5)$$

where the permutation is solved before calculating SIR and SDR, i.e., y_i is the estimation of s_i, and y_{is_j} is the output of the whole separating system at y_i when only s_j is active and $s_k (k \neq j)$ does not exist, and x_{ks_j} is the observation obtained by microphone k when only s_j exists. α and D are parameters to compensate for the amplitude and phase difference between x_{ks_i} and y_{is_i}. To evaluate the conventional method (sparseness only method), we calculated SIR and SDR using the measurements from both microphones, and adopted the better value.

5.3. Experimental results

Figure 4 shows the power content by percentage of each signal when $\Delta = 10°$. From the upper three rows of Fig. 4, we can see that the observations in $\tilde{\theta_2} - \Delta \leq \theta(\omega, m) \leq 180°$ mainly contain the signals $\tilde{S_2}$ and $\tilde{S_3}$, on the other hand, the observations in $0° \leq \theta(\omega, m) \leq \tilde{\theta_2} + \Delta$ mainly contain the signals $\tilde{S_1}$ and $\tilde{S_2}$. From this, we set $\theta_{th1} = \tilde{\theta_2} - \Delta$ for $\tilde{S_1}$ removal (case 1), and $\theta_{th2} = \tilde{\theta_2} + \Delta$ for $\tilde{S_3}$ removal (case 2). In Fig. 4, the percentages of each power

Table 1: Power lost by binary masks (in %)

mask	M_1	M_2	M_3	M_{ICA}^{12}		M_{ICA}^{23}	
output	Y_1	Y_2	Y_3	Y_1	Y_2	Y_2	Y_3
[%]	17	14	23	2.5	5.7	8.1	0.7

extracted by the binary mask (2) with these thresholds are also shown. Because the proportion of the *third* signal is very small, we can say that we can use an ICA at the 2nd stage of our method.

Table 1 shows the signal power eliminated by the zero-padding $\frac{\sum_n s_i(n)^2 - \sum_n \hat{s}_i(n)^2}{\sum_n s_i(n)^2}$ caused by binary masks, where $\hat{s_i}(n) = \mathrm{IDFT}[M_i(\omega, m)S_i(\omega, m)]$. If we use these mask thresholds for separation, i.e., in the sparseness only case, around 20% of signal power was diminished by the binary mask. By contrast, with our method, the signal power eliminated by $M_{\mathrm{ICA}}^{pq}(\omega, m)$ was inferior. This result convinces us that the adverse effect of zero-padding was mitigated by using our method, that is, one source removal.

Table 2 shows the experimental results of anechoic simulations. The first row shows the results obtained solely using the sparseness, the second and third rows show the results obtained with our SPICA method. With sparseness only, the SIR values were high but the SDR values were unsatisfactory. A large musical noise was heard in this case. In contrast, with SPICA, we were able to obtain high SDR values without any serious deterioration in the separation performance SIR. Now the musical noise decreased.

Moreover, we conducted reverberant tests (T_R=130 ms) with settings of 50° (female), 100° (male), 135° (male), and 45° (female), 90° (male), 135° (male). The results are shown in Table 3. Even in a reverberant environment, we obtained reasonable results with our proposed SPICA.

It should be noted that it is hard to separate the signal at the center position by both methods.

Some sound samples can be found at our web site [10].

6. Conclusion

We proposed a method for separating more speech signals than sensors by combining a sparseness approach and ICA (SPICA). Our method avoids over-zero-padding, and therefore, can separate the signals with less distortion in a reverberation of 130 ms.

7. References

[1] S. Rickard and O. Yilmaz, "On the W-Disjoint orthogonality of speech," *Proc. ICASSP2002*, vol.1, pp. 529-532, 2002.

[2] L. Vielva, D. Erdogmus, C. Pantaleon, I. Santamaria, J. Pereda and J. C. Principe, "Underdetermined blind source separation in a time-varying environment," *Proc. ICASSP2002*, vol. 3, pp. 3049-3052, 2002.

[3] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," *Proc. ICA2000*, pp. 87-92, 2000.

[4] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T, Sakurai and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoust. Sci. & Tech*, vol. 22, no. 2, pp. 149-157, 2001.

[5] A. Blin, S. Araki and S. Makino, "Blind source separation when speech signals outnumber sensors using a sparseness-mixing matrix combination," *Proc. IWAENC2003*, 2003.

[6] S. Araki, R. Mukai, S. Makino, T. Nishikawa and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech,"*IEEE Trans. on SAP* vol. 11, no. 2, pp. 109-116, 2003.

[7] H. Sawada, R. Mukai, S. Araki and S. Makino, "Polar coordinate based nonlinear function for frequency domain blind source separation," *Proc. ICASSP2002*, pp. 1001-1004, May 2002.

[8] H. Sawada, R. Mukai, S. Araki, S. Makino, "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation," *Proc. ICA2003* pp. 505-510, Apr. 2003.

[9] K. Matsuoka and S. Nakashima, "A robust algorithm for blind separation of convolutive mixture of sources," *Proc. ICA2003*, pp. 927-932, Apr. 2003.

[10] http://www.kecl.ntt.co.jp/icl/signal/araki/underdeterminedBSSdemo.html

Table 2: Results of anechoic simulations. Signals are at $50^\circ, 100^\circ, 135^\circ$. Sparse: with only sparseness, Case 1: with SPICA ($\tilde{S}_1$ was removed for ICA), Case 2: with SPICA ($\tilde{S}_3$ was removed for ICA).

female-male-male						[dB]
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Sparse	17.6	11.6	17.3	7.3	9.3	8.5
Case 1	-	8.8	13.6	-	12.5	**16.2**
Case 2	17.5	8.8	-	**20.8**	11.8	-
male-male-male						
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Sparse	13.1	7.7	15.6	4.3	8.1	4.6
Case 1	-	4.5	10.4	-	9.6	**10.0**
Case 2	12.6	4.0	-	**17.7**	13.5	-
female-female-female						
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Sparse	23.6	11.3	18.2	8.5	11.9	8.5
Case 1	-	8.0	13.0	-	15.3	**15.6**
Case 2	16.5	8.0	-	**21.5**	14.2	-

Table 3: Results of reverberant tests (T_R=130ms) for a female-male-male combination. Sparse: with only sparseness, Case 1: with SPICA ($\tilde{S}_1$ was removed for ICA), Case 2: with SPICA ($\tilde{S}_3$ was removed for ICA).

$50^\circ, 100^\circ, 135^\circ$						[dB]
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Sparse	9.9	4.7	8.3	4.0	8.3	4.8
Case 1	-	4.1	9.6	-	9.1	**6.9**
Case 2	9.0	3.3	-	**9.4**	10.3	-
$45^\circ, 90^\circ, 135^\circ$						
	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
Sparse	12.1	5.8	13.8	4.3	9.7	3.5
Case 1	-	5.3	12.1	-	10.9	**8.4**
Case 2	8.4	4.2	-	**7.6**	11.5	-