Blind Source Separation when Speech Signals Outnumber Sensors using a Sparseness - Mixing Matrix Estimation (SMME)

Audrey Blin *Shoko Araki* *Shoji Makino*
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, JAPAN
e-mail: blin@cslab.kecl.ntt.co.jp

Abstract

This paper focuses on underdetermined blind source separation (BSS) of three speech signals mixed in a real environment from measurements provided by two sensors. Underdetermined BSS is a problem that has not yet been intensely studied and so far no satisfying solution has been obtained. The major issue encountered in previous work relates to the occurrence of distortion, which affects a separated signal with loud musical noise. To overcome this problem, we propose combining sparseness with the use of an estimated mixing matrix. First, we use a geometrical approach to perform a preliminary separation and to detect when only one source is active. This information is then used to estimate the mixing matrix, which allows us to improve our separation. Experimental results show that this Sparseness - Mixing Matrix Estimation (SMME) provides separated signals of better quality (less distortion, less musical noise) than those extracted without using the estimated mixing matrix.

1. Introduction

BSS refers to the problem of recovering signals from several observed linear mixtures. The adjective “blind” stresses the fact that the source signals are not observed and that no information is available about how the sources are mixed.

In a determined problem a common way of solving the BSS issue was to use Independent Component Analysis (ICA) [1], or to estimate and then invert the mixing matrix modeling the system. But, here, where sources outnumber sensors, the mixing matrix is no longer square and the previous solutions are not feasible. In fact, more information is needed to solve an underdetermined problem.

Up to now, solving the BSS problem in an underdetermined case has mainly consisted in assuming that the speech signals were sufficiently sparse [2-6]. However, due to unexpected discontinuous zero-padding, such separated signals have considerable distortion, and therefore a loud musical noise is heard.

In [5], an estimated mixing matrix was used for solving the determined BSS problem. Our suggestion for eliminating the distortion matter is to combine sparseness with a mixing matrix estimation. Indeed we can obtain more information about the signals to be separated and hence reduce the zero-padding effect, from which the musical noise originates.

Whereas Vielva et al., Rickard and Yilmaz worked on an underdetermined instantaneous case employing sparseness [2,3,4], and Deville on a determined instantaneous case utilizing a mixing matrix estimation [5], here, we are dealing with underdetermined BSS in a convolutive case.

In this paper, essential investigations with regard to sparseness are described in Sec. 3. Section 4 is a step-by-step presentation of the proposed method, whose results are provided in Sec. 5. The paper concludes in Sec. 6.

2. Problem statements and notations

In this paper, we consider speech mixtures observed in a real room. In this case, as speeches are mixed with their reverberation, the observed vectors x_j ($j=1..M$) can be modeled as convolutive mixtures of the source signals s_i ($i=1..N$) as follows:

$$x_j(t)=\sum_{i=1}^{N} h_{ji} * s_i(t) \qquad (1)$$

where h_{ji} is the impulse response from a source i to a sensor j. In this paper, we deal with a case where $N=3$ sources and $M=2$ sensors. Moreover, we assume that the source signals are mutually independent and sparse: namely signals have large values at rare sampling points. We are using the Short Time Fourier Transform (STFT) to convert our problem into a linear instantaneous mixtures' problem as well as to improve the sparseness of the speech signals [4]. In the time-frequency domain, our system becomes: $\boldsymbol{X}(f,m)=\boldsymbol{H}(f)\boldsymbol{S}(f,m)$ where f is the frequency, m the frame index, $\boldsymbol{H}(f)$ the 2×3 mixing matrix whose i-j component is a transfer function from a source i to a sensor j, $\boldsymbol{X}(f,m)=[X_1(f,m), X_2(f,m)]^T$ and $\boldsymbol{S}(f,m)=[S_1(f,m), S_2(f,m), S_3(f,m)]^T$, namely the Fourier transformed observed signals and source signals, respectively.

Our aim is to estimate three speech signals from measurements provided by two sensors.

3. Sparseness Inquiries

The first definition of sparseness is that the more zero samples contained in a source, the more sparse it is, which means that the sources overlap at infrequent intervals.

Figure 1 is a histogram showing the number of sources that are simultaneously active. It can be seen that the time points where no sources are active are very numerous whereas the time points where three sources are active are very infrequent. We can infer from these observations that the signals are sparse and that the three signals rarely overlap.

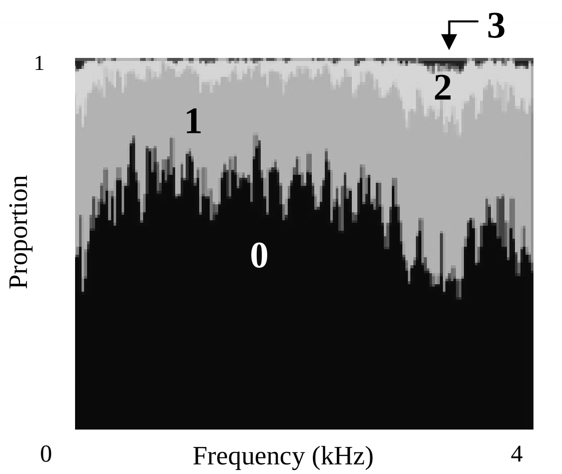

Fig. 1: Histogram of the number of active sources: 0, 1, 2 or 3 for a male-male-female combination recorded with a reverberation of 200 ms and for an fftsize of 512.

3.1. Measure of overlapping

In this section, we investigated the sparseness more closely checked the degree of signal overlap by utilizing a criterion called Approximate W-Disjoint Orthogonality (WDO) defined by Rickard and Yilmaz [6].

We use a mask:

$$\phi_{(j,x)}(f,m) = \begin{cases} 1 & 20\log(|S_j(f,m)|/|Y_j(f,m)|) > x \\ 0 & otherwise \end{cases} \quad (2)$$

where $Y_j(f,m)$ is the STFT of $y_j(t) = \sum_{i=1, i\neq j}^{N} s_i(t)$ i.e. $y_j(t)$ is the summation of the sources interfering with source j. The Approximate WDO is defined as:

$$r_j(x) = 100 \,\|\phi_{(j,x)}(f,m) S_j(f,m)\|^2 \,/\, \|S_j(f,m)\|^2 \quad (3)$$

This measures the percentage r_j of source j energy for time-frequency points where it dominates the other signals by r_j % at x dB. From this criterion it emerges that, if we can predict the time-frequency points at which a source dominates the others by r_j % at x dB, we should be able to recover r_j % of the energy of the original sources. If r_j is sufficiently large, we can separate signals with small distortion and vice-versa.

For example in Fig. 2, if we want a signal-to-interference ratio of 20 dB, only around 50 % of the original power is recoverable, which means that almost half the points are zero-padded by a mask and such distortion cannot be avoided.

Moreover Fig. 2 shows that reverberant data have a lower Approximate WDO than unreverberant data. Hence separating reverberant data becomes more difficult.

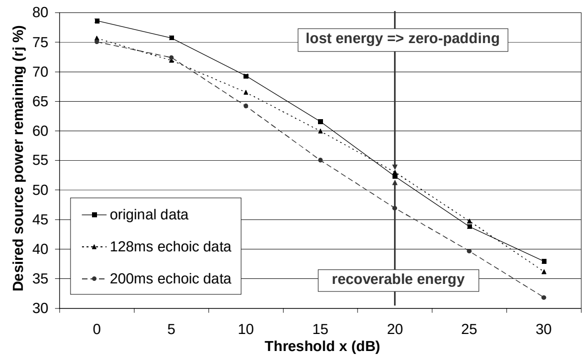

Fig. 2: Approximate WDO against the threshold x for an fftsize of 512 and a male-male-female combination.

3.2. The sparsest representation with STFT

To make the separation easier, we have to represent our data in the domain where their degree of sparseness is the best. By plotting the Approximate WDO against the fftsize in Fig.3, we can estimate the appropriate fftsize for the easiest separation. Hence, from now, we will use an fftsize of 512 for a 8 kHz sampling rate in as much as this fftsize allows the best recovery in terms of energy and therefore separated signals are less distorted.

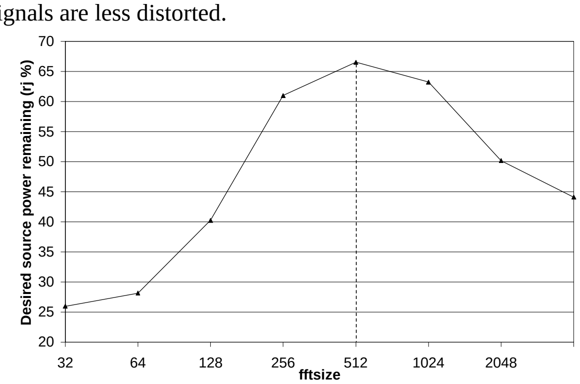

Fig.3: Approximate WDO of anechoic measured data against the fftsize for a threshold x of 10 dB and a sampling rate of 8 kHz.

4. Proposed method: Sparseness - Mixing Matrix Estimation (SMME)

In previously reported methods [2-4], one of the major drawbacks was the occurrence of distortion, i.e., musical noise. To overcome this issue, we propose a three-step method. First, using the sparseness of speech signals, we adopt a geometrical approach extracting the time points m when only one source is active [**1rst step**], then we estimate the mixing matrix [**2nd step**] and finally we reconstruct the signals when two sources are active [**3rd step**].

- **[1rst step]** ***Geometrical approach***

This first step consists of detecting the frame indices m when only one of the three sources is active for each frequency bin f.

Scatter-plots of the measurements, as shown in Fig. 4, comprise three main lines (if the sources are sparse enough).

According to Vielva et al. [4], these lines symbolize the directions defined by the column vectors of the mixing matrix. In other words, they can be seen as a representation of each source existing alone. In between two given directions, we find the time-frequency points modeling our system when two sources (those linked to the above directions) are active simultaneously.

By setting narrow areas each containing only one line, such as areas 1, 2 and 3 in Fig. 4, we are able to determine when only one source is active. At the same time we can also reconstruct the signals for these time-frequency points. This is the method exploited in previous works [2, 3, 4]. However, as expected by using such a rough approach, the quality of the separated signals is not satisfactory. Since the rate of recoverable energy is too low (as shown in Fig. 6), we cannot avoid an important zero-padding, which makes the signals insufficiently continuous. As a result, considerable distortion i.e., loud musical noise can be heard.

To overcome this lack of quality, we attempt to complete our separation using a totally different approach, relying on the knowledge of the mixing matrix.

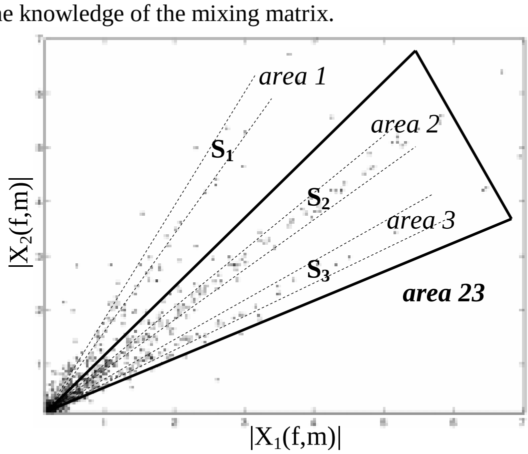

Fig. 4: Scatter-plots of the mixtures at a frequency of 312 Hz for male-male-female combination, a reverberation of 130 ms and an fftsize of 512.

- **[2nd step]** ***Estimation of mixing matrix***

Deville recovers the mixing matrix by estimating a certain cross-correlation parameter ratio over time-frequency zones where only one source exists [5]. This ratio was then proved to be equal to H_{2i}/H_{1i} ($i=1, 2, 3$).

In constrast to Deville, here we are working with a underdetermined convolutive case, however his approach gave us the idea to model our system in the time-frequency domain by:

$$\begin{bmatrix} X_1(f,m) \\ X_2(f,m) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ \frac{H_{21}(f)}{H_{11}(f)} & \frac{H_{22}(f)}{H_{12}(f)} & \frac{H_{23}(f)}{H_{13}(f)} \end{bmatrix} \cdot \begin{bmatrix} H_{11}(f)S_1(f,m) \\ H_{12}(f)S_2(f,m) \\ H_{13}(f)S_3(f,m) \end{bmatrix} \quad (4)$$

Therefore, using time-frequency points estimated in the first step when only S_i ($i=1, 2, 3$) is active, we have:

$$\begin{cases} X_1(f,m) = H_{1i}(f)S_i(f,m) \\ X_2(f,m) = H_{2i}(f)S_i(f,m) \end{cases} \quad (5)$$

whose ratio $X_2(f,m)/X_1(f,m)$ provides one of the components of the mixing matrix $H_{2i}(f)/H_{1i}(f)$.

- **[3rd step]** ***Reconstruction of time-frequency points when two sources are active***

At this stage, it should be noted that knowing the mixing matrix does not enable us to separate the signals when three sources are active. This is because the mixing matrix is not square and does not have any inverse. Deville has only applied his method to a squared mixing matrix. Nevertheless, it is still possible to rebuild the time-frequency points when two sources are active, providing that for each frequency bin, we know the frame indices for which this case occurs. Once more this information is provided by the geometrical approach employed in the first step. But this time, instead of setting the limits very close to the observed directions, we are considering much wider areas so as to enclose the points located between two given directions. Indeed let us suppose that, for an estimated *(f, m)* detected during the first step, $S_1(f, m)$ is null (area 23 in Fig. 4), in this area, our system becomes:

$$\begin{bmatrix} X_1(f,m) \\ X_2(f,m) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \frac{H_{22}(f)}{H_{12}(f)} & \frac{H_{23}(f)}{H_{13}(f)} \end{bmatrix} \cdot \begin{bmatrix} H_{12}(f)S_2(f,m) \\ H_{13}(f)S_3(f,m) \end{bmatrix} \quad (6)$$

Now the mixing matrix is square and can thus be inverted, leading to $H_{12}(f)S_2(f,m)$ and $H_{13}(f)S_3(f,m)$:

$$\begin{bmatrix} H_{12}(f)S_2(f,m) \\ H_{13}(f)S_3(f,m) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \frac{H_{22}(f)}{H_{12}(f)} & \frac{H_{23}(f)}{H_{13}(f)} \end{bmatrix}^{-1} \cdot \begin{bmatrix} X_1(f,m) \\ X_2(f,m) \end{bmatrix} \quad (7)$$

Moreover, in this area, if the signals $H_{12}(f)S_2(f,m)$ and $H_{13}(f)S_3(f,m)$ are not too greatly zero-padded, we expect that the distortion of the estimated $H_{12}(f)S_2(f,m)$ and $H_{13}(f)S_3(f,m)$ will not be that large. We proceed in the same way when $S_3(f,m)$ is null.

It should be noted that, in Fig. 1, we have already confirmed that we do not often have three sources active simultaneously.

5. Experiments

5.1. Experimental conditions

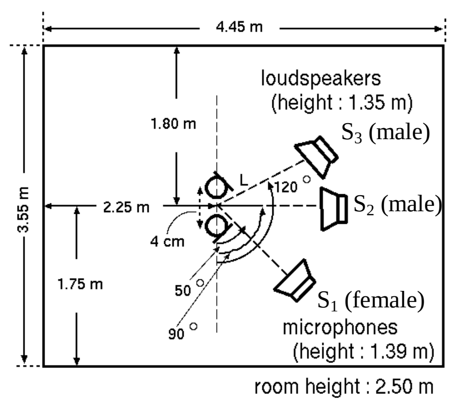

Fig. 5: Experimental conditions

The recordings were done in a room with little reverberance (T_R=130 ms) using a two-element array of directional microphones 4 cm apart. The speech signals, sampled at 8 kHz, came from three directions: 120° (male), 90° (male) and 50° (female) and the distance between the sources and the sensors was L = 55 cm.

5.2. Mask justification

Figure 6 justifies our decision to use wide masks. Indeed if we use narrow masks (e.g., area 3 in Fig. 4) as in the previous method, the recoverable power is only around 45 % with a threshold of 10 dB whereas if we utilize wider masks (e.g., area 23 in Fig. 4), we can recover over 60 % of this power.

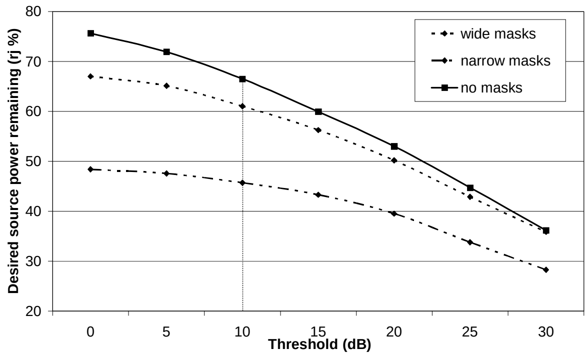

Fig. 6: Approximate WDO against the threshold, fftsize=512, T_R=130 ms.

Consequently the technique using wide areas makes it possible to reduce the distortion of the separated signals, which was our aim.

5.3. SIR and SDR calculations

To evaluate the separation performance of our method, we have chosen to calculate 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:

$$\mathrm{SIR_i} = 10\log\frac{\sum_t y_{is_i}^{2}(t)}{\sum_{i\neq j}\sum_t y_{is_j}^{2}(t)} \quad (8)$$

$$\mathrm{SDR_i} = 10\log\frac{\sum_t x_{ks_i}^{2}(t)}{\sum_t (x_{ks_i}(t) - \alpha y_{is_i}(t-\varphi))^2} \quad (9)$$

where the permutation is solved before calculating SIR and SDR, i.e. $y_i(t)$ is the estimation of $s_i(t)$, and y_{iSj} 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_{kSj} is the observation obtained by microphone k when only s_j exists. α is a constant that compensates for the amplitude difference and φ is an angle that fits the phase difference between input x_{kSj} and output y_{iSj}. To evaluate the performance of the previous method (sparseness only method), we calculated SIR and SDR using both microphones' measurements, and adopted the better values.

5.4. Results

Table 1 shows the results we obtained from our measurements. By "sparseness" we imply that we are evaluating the performance of our speech signals when we are applying the narrow masks. "invH12" means that we are applying our mixing matrix to area 12 comprising speech signals 1 and 2. Likewise "invH23" means that we are applying our mixing matrix to area 23 comprising speech signals 2 and 3. Actually, we are comparing the conventional method with our SMME method.

Table 1: SIR and SDR calculated in dB for different approaches, fftsize=512, T_R=130 ms

	SIR1	SIR2	SIR3	SDR1	SDR2	SDR3
sparseness	15.3	9.9	10.6	8.4	10.3	3.4
invH12	11.6	3.1		**8.7**	12.2	
invH23		3.3	7.6		12.5	**7.2**

As we can see, the use of our SMME method allows us to obtain less distorted signals.

We performed informal listening tests and it is important to note that, for the third signal, although the two SDR results are very close, much less musical noise is heard when separation is undertaken using SMME than when only sparseness is used.

6. Conclusion

We proposed a separation method for use when there are more speech signals than sensors by combining a sparseness approach and an estimation of the mixing matrix. The first experimental results are very encouraging in terms of quality and suggest that the SMME is an approach that deserves serious investigation.

7. References

[1] A. Hyvarinen, J. Karhunen and E. Oja, "Independent Component Analysis," John Wiley & Sons, 2001.

[2] 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.

[3] M. Zibulesky, B.A. Pearlmutter, P. Bofill and P. Kisilev, "Blind source separation by sparse decomposition in a signal dictionary," TR No. CS99-1, University of New Mexico, Albuquerque, July 1999.

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

[5] Y. Deville, "Temporal and time frequency correlation-based blind source separation methods," Proc. ICA2003, pp. 1059-1064, April 2003.

[6] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," Proc. ICASSP2002, vol.1, pp. 529-532, 2002.