PAPER Special Section on Multi-channel Acoustic Signal Processing

Underdetermined Blind Separation of Convolutive Mixtures of Speech Using Time-Frequency Mask and Mixing Matrix Estimation

Audrey BLIN[†*a)], Shoko ARAKI[†], *Nonmembers, and* Shoji MAKINO[†], *Member*

SUMMARY This paper focuses on the underdetermined blind source separation (BSS) of three speech signals mixed in a real environment from measurements provided by two sensors. To date, solutions to the underdetermined BSS problem have mainly been based on the assumption that the speech signals are sufficiently sparse. They involve designing binary masks that extract signals at time-frequency points where only one signal was assumed to exist. 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 detect when only one source is active and to perform a preliminary separation with a time-frequency mask. This information is then used to estimate the mixing matrix, which allows us to improve our separation. Experimental results show that this combination of time-frequency mask and mixing matrix estimation provides separated signals of better quality (less distortion, less musical noise) than those extracted without using the estimated mixing matrix in reverberant conditions where the reverberant time (TR) was 130 ms and 200 ms. Furthermore, informal listening tests clearly show that musical noise is deeply lowered by the proposed method comparatively to the classical approaches.

key words: *source separation, blind, underdetermined, convolutive, sparseness of speech, distortion, time-frequency mask, mixing matrix, musical noise*

1. Introduction

The title of this paper includes the words blind, underdetermined and convolutive and each should be paid a particular attention. They are key words with regard to our work and each of them reflects a level of difficulty we have to face. Therefore, in this introduction, we will focus on each in turn and explain the way they relate to each other.

First, 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 (e.g., [1]). Blind source separation (BSS) actually refers to the problem of recovering signals from several observed linear mixtures. The weakness of the prior information is precisely the strength of the BSS model.

In this paper, we concentrate on the audio BSS issue, which is currently a major research field. Probably the best-known problem in auditory scene analysis is that of the cocktail party.

Recently, many methods have been proposed to solve the BSS problem of audio signals in real environments. A common approach involves using independent component analysis (ICA) (e.g., [2], [3]). However, most of these methods consider a determined problem, where we have as many sources as sensors (e.g. [4]–[6]).

Here, in an attempt to deal with a more contemporary issue, we have decided to concentrate on an underdetermined case where there are more sources than sensors. This choice comes from the fact that, in a real situation, most of the time, there are ambient and uncountable noise sources. Furthermore, manufacturers do not want many sensors for their small and inexpensive products. Therefore, for real applications, we have to consider the underdetermined case.

Moreover, to be more realistic, we should consider the mixing process of the speech signal to be convolutive because most speech signals are recorded with their reverberation. Below we focus on the two remaining key words and look at the kinds of solutions that have already been proposed for dealing with the issues that the words refer to.

The true obstacle raised by underdetermination relates to the fact that a simple inversion of the mixing matrix will not lead us to a solution in as much as the mixing system is not square and therefore not invertible. Furthermore, even if we knew the mixing matrix exactly, we would not be able to recover the original signals because part of the information is lost during the mixing process.

Thus far, solving the BSS problem in an underdetermined case has mainly consisted in assuming that the speech signals were sufficiently sparse [7]–[11], which legitimizes the extraction of each source.

It is our understanding that there are two approaches relying on sparseness that will solve the underdetermined BSS. The first involves the clustering of time-frequency points with binary masks [10]. It emerges that, if the signals are sufficiently sparse, namely if most of the samples of a signal are almost zero, we can assume that the sources rarely overlap. [10] uses this assumption and extracts each signal using a time-frequency binary mask. The second approach is based on ML estimation, where the sources are estimated by l_1-norm minimization after an evaluation of the mixing matrix [7], [9], [12], [13].

Although, dealing with the underdetermined BSS is already very tough, it is still possible to present ourselves with a greater challenge by considering the underdetermined *convolutive* BSS issue, a solution to which has already been

Manuscript received November 2, 2004.
Manuscript revised January 21, 2005.
Final manuscript received March 11, 2005.

[†] The authors are with the NTT Communication Science Laboratories, NTT Corporation, Kyoto-fu, 619-0237 Japan.

[*] Presently, with the Université du Québec, INRS EMT, 800 de la Gauchetière Ouest, Suite 6900, Montréal, Québec, H5A1K6, Canada.

a) E-mail: audrey_blin@hotmail.com

DOI: 10.1093/ietfec/e88–a.7.1693

sought [10]. However, this approach leads to the discontinuous zero-padding of separated signals because all the unknown samples (e.g. when several sources are active at the same time) are set at zero by a binary mask. Consequently such separated signals are greatly distorted, and therefore a loud musical noise is heard. In contrast to the ML approach, which has excited great interest in recent years, only a few trials have been undertaken related to the convolutive case. Of course, convolutive mixtures can be seen frequency bin by frequency bin as instantaneous mixtures but, as such a conversion results in complex-valued signals, the l_1-norm minimization is not applicable straightforwardly and would need to be expanded or alternated. Consequently, due to a high level of complexity, the ML approach is rather avoided in the convolutive case.

The objective of our work is to provide a satisfactory solution to the underdetermined convolutive BSS by ensuring that any musical noise is minimized. In this paper, we focus on the underdetermined BSS of three speech signals mixed in a real environment from measurements provided by two sensors.

It should be remembered that one of the ways to overcome the BSS issue in a determined problem is to estimate and then invert the mixing matrix modelling of our system [8]. Recently, [14] proposed an approach for estimating the separating matrix in the instantaneous determined BSS problem. However, here, where sources outnumber sensors, the mixing matrix and separating matrix are no longer square and such solutions cannot be used. Nevertheless, this gave us the idea of combining the sparseness properties of speech signals with an estimation of the mixing matrix.

Our suggestion for eliminating the distortion problem in convolutive underdetermined BSS is to combine sparseness with a mixing matrix estimation [15]. First, using the sparseness, we detect the time points when only one source is active. This information is then used to estimate the mixing matrix. Subsequently, by using sparseness investigations, we were able to work under conditions where one of the sources could be omitted by a time-frequency mask. The separation of the residual signals using the inverse of the estimated mixing matrix then follows. By this one source removal, the zero-padding of the separated signal is expected to be less corrupting than with the previous binary mask approach. Results and informal listening show that we indeed obtain more information about the signals to be separated and significantly lowered the musical noise comparatively to the classical approaches. As a consequence, we can reduce the zero-padding effect from which the musical noise originates. A complementary approach, also relying on one source removal, has been proposed by the same authors [16], where they use the ICA algorithm to the residual signals. Apart from Araki's approach and so as to overcome the previously mentioned lack of quality, here, we attempt to utilize an estimated mixing matrix to separate the residual.

The organization of this paper is as follows. Section 2 describes our problem statements and essential investigations with regard to sparseness are described in Sect. 3. Section 4 is a step-by-step presentation of the proposed method, whose results are provided in Sect. 5. Section 6 provides our conclusions.

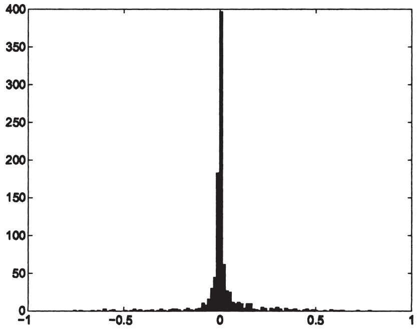


Fig. 1 An example in the frequency domain (250 Hz) of a probability density function of a speech signal.

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 $(j = 1, \cdots, M)$ can be modelled as convolutive mixtures of the source signals s_i $(i = 1, \cdots, N)$ as follows:

$$x_j(n) = \sum_{i=1}^{N} \sum_{l} h_{ji}(l) s_i(n - l + 1) \tag{1}$$

where h_{ji} is the impulse response from a source i to a sensor j. As it is our first trial, we decided to deal with the simplest case, i.e. where $N = 3$ sources and $M = 2$ sensors. Moreover, we assume that the source signals are sparse: namely signals have large values at rare sampling points as shown by the probability density function (pdf) of a speech signal in Fig. 1.

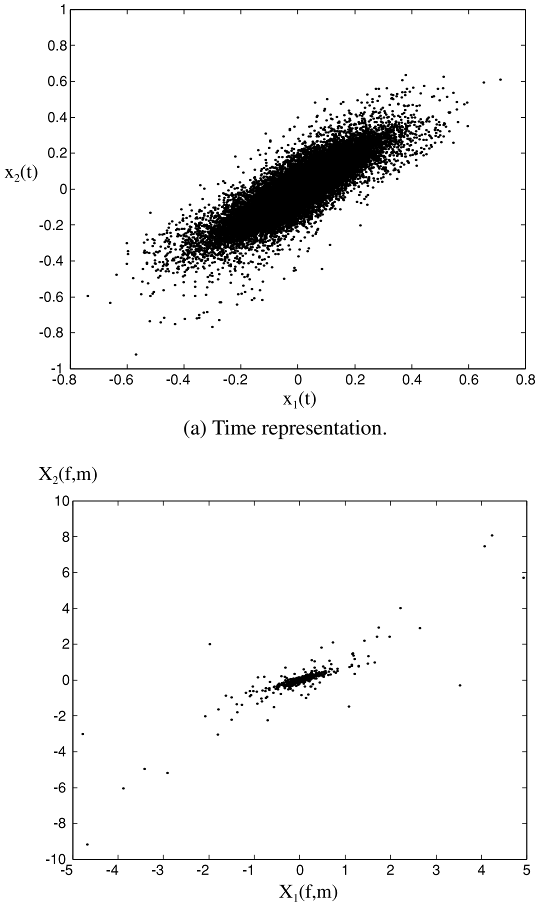
We use the Short Time Fourier Transform (STFT) to convert our problem into a linear instantaneous mixture problem as well as to improve the sparseness of the speech signals [9] as shown in Fig. 2. In the time-frequency domain, our system becomes:

$$\mathbf{X}(f, m) = \mathbf{H}(f)\mathbf{S}(f, m) \tag{2}$$

where f is the frequency, m the frame index, $\mathbf{H}(f)$ the 2×3 mixing matrix whose (j, i) component is a transfer function from a source i to a sensor j, $\mathbf{X}(f, m) = [X_1(f, m), X_2(f, m)]^T$ and $\mathbf{S}(f, m) = [S_1(f, m), S_2(f, m), S_3(f, m)]^T$, 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

Sparseness becomes greater as the number of zero samples contained in a source increases, which means that the sources overlap at infrequent intervals. Interesting investigations of the sparseness property of speech signals have



(a) Time representation.

(b) Frequency representation (f=1078 Hz).

Fig. 2 Scatter plot of X_2 versus X_1 of three mixed speech for a male-male-female combination.

been proposed in [10] for anechoic speech signals and in [17] for two speech signals. Here, we give detailed investigations for three speech signals not only for anechoic signals but also echoic signals.

Figure 3 is a histogram showing the number of sources that are simultaneously active. In the figure, we consider the signal $s_i(n)$ is active when the signal $s_i(n)$ has a greater amplitude than $\max(|s_k(n)|)/10$ ($k = 1, 2, 3, n = wholelength$). It can be seen that there are many time points at which no sources are active and few where three sources are active. We can infer from these observations that the signals are sparse and that three signals rarely overlap.

3.1 Measure of Overlapping

To determine the best representation, we investigated the sparseness more closely and checked the degree of signal overlap by utilizing a criterion called Approximate W-Disjoint Orthogonality (WDO) defined by Yilmaz and Rickard [10]. We use a mask:

$$\Phi_{(j,x)}(f,m) = \begin{cases} 1, & \text{if } 20\log\left(\frac{|S_j(f,m)|}{|Y_j(f,m)|}\right) > x \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

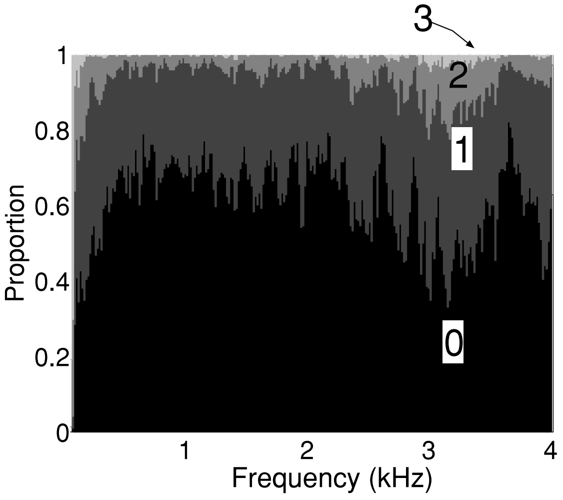


Fig. 3 Histogram of the number of active sources: 0, 1, 2 or 3 for a male-male-female combination recorded with a reverberation of 200 ms and for a DFTsize of 512.

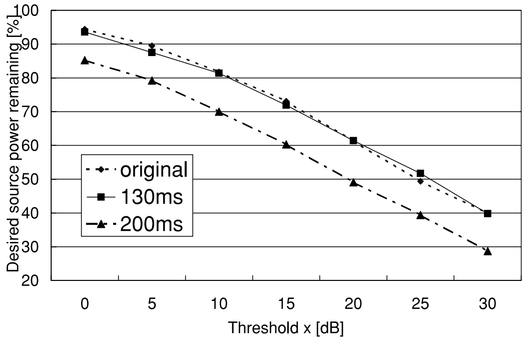


Fig. 4 Approximate WDO against the threshold x for a DFTsize of 512, a male-male-female combination using a reverberant time of 130 ms or 200 ms, which can be compared to the original recording.

where $Y_j(f,m)$ is the DFT of

$$y_j(n) = \sum_{i=1, i\neq j}^{N} s_i(n) \tag{4}$$

i.e. $y_j(n)$ is the summation of the sources interfering with source j. The approximate WDO is defined as:

$$r_j(x) = 100\frac{\|\Phi_{(j,x)}(f,m)S_j(f,m)\|^2}{\|S_j(f,m)\|^2}, \tag{5}$$

where $\|f(x,y)\|^2 = \sum_y \sum_x |f(x,y)|^2$, where f is a function of x and y.

This measures the percentage r_j of source j energy for time-frequency points where this source 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 little distortion and vice-versa. For example in Fig. 4, if we want a signal-to-interference ratio of 20 dB, only around 60% of the

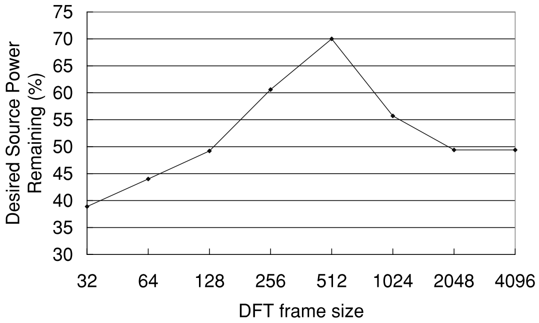


Fig. 5 Approximate WDO of anechoic measured data against the DFT-size for a threshold x of 10 dB and a sampling rate of 8 kHz, TR = 200 ms.

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. 4 shows that reverberant data have a lower Approximate WDO than anechoic data. Hence separating reverberant data becomes more difficult.

3.2 Sparsest Representation with STFT

To make the separation easier, we have to represent our data in the space domain where their degree of sparseness is the best. Figure 5 shows the Approximate WDO against the different frame size for STFT, which called DFTsize (Discrete Fourier Transform frame size) in this paper. From Fig. 5, we can estimate the appropriate DFTsize for the easiest separation. Hence, from now, we will use a DFTsize of 512 for a 8 kHz sampling rate in as much as this DFTsize allows the best recovery in terms of energy and therefore separated signals are less distorted.

4. Proposed Method

With previously reported methods [7], [9], [11], a major drawbacks of the classical approaches was the occurrence of distortion, i.e., musical noise. To overcome this problem, we propose a three-step method. First, using the sparseness of speech signals, we adopt a geometrical approach to determine the time-frequency masks that extract the time points m when only one source is active [1st step], then we estimate the mixing matrix [2nd step] and finally we reconstruct the signals when two sources are active [3rd step].

• [1st **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. 6, comprise three main lines (if the sources are sparse enough). According to Vielva et al. [9], these lines represent 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 modelling our system when two sources (those linked to the above directions) are active simultaneously.

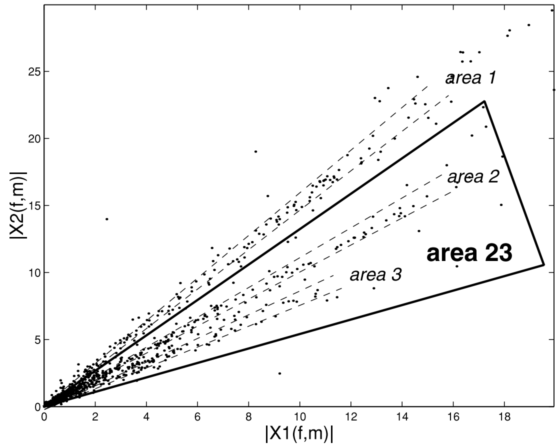


Fig. 6 Scatter-plots of the mixtures at a frequency of 312 Hz for a male-male-female combination, a reverberation of 130 ms and a DFTsize of 512.

By setting narrow areas each containing only one line, such as areas 1, 2 and 3 in Fig. 6, we are able to determine when only one source is active and at the same time we can reconstruct the signals for these time-frequency points. That is to say that three time-frequency binary masks can be designed, they are assigned the value 1 in areas 1, 2 or 3 and the value 0 otherwise. This is basically the method exploited in previous work [10]. However, as expected when using such a rough approach, the quality of the separated signals is unsatisfactory. Since the rate of recoverable energy is very low (as shown in Fig. 9), we cannot avoid an important zero-padding, which makes the signals insufficiently continuous. As a result, we hear considerable distortion i.e., loud musical noise. To recover this lack of quality, we attempt to complete our separation relying on the knowledge of the mixing matrix. Distinctly from Araki's approach [16], we try to utilize the estimated mixing matrix to separate the residual.

• [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 [8]. This ratio was then proved to be equal to H_{2i}/H_{1i} $(i = 1, 2, 3)$. In contrast to Deville, here we are working with an underdetermined convolutive case, however his approach gave us the idea of modelling our system in the time-frequency domain by:

$$\begin{pmatrix} X_1(f,m) \\ X_2(f,m) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ \frac{H_{21}}{H_{11}} & \frac{H_{22}}{H_{12}} & \frac{H_{23}}{H_{13}} \end{pmatrix} \begin{pmatrix} H_{11}(f)S_1(f,m) \\ H_{12}(f)S_2(f,m) \\ H_{13}(f)S_3(f,m) \end{pmatrix} \tag{6}$$

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

$$\begin{cases} X_1(f,m) = H_{1i}(f)S_i(f,m) \\ X_2(f,m) = H_{2i}(f)S_i(f,m) \end{cases} \tag{7}$$

whose ratio $X_2(f,m)/X_1(f,m)$ provides one of the components of the mixing matrix $H_{2i}(f)/H_{1i}(f)$. For a stable estimation of the mixing matrix coefficients, we estimated the expectation of the ratio X_2/X_1:

$$H_{2i}(f)/H_{1i}(f) = \mathbf{E}[X_2(f,m)/X_1(f,m)] \tag{8}$$

where **E[.]** is the expectation at time m when $X_1(f,m)$ and $X_2(f,m)$ are in area i.

- [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 an inverse. Deville [8] 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. That is to say that we utilize wider time-frequency masks here. Indeed let us suppose that, for an estimated (f,m) detected during the first step, $S_1(f,m)$ is zero (area 23 in Fig. 6), in this area, our system becomes:

$$\begin{pmatrix} X_1(f,m) \\ X_2(f,m) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \frac{H_{22}}{H_{12}} & \frac{H_{23}}{H_{13}} \end{pmatrix} \begin{pmatrix} H_{12}(f)S_2(f,m) \\ H_{13}(f)S_3(f,m) \end{pmatrix} \tag{9}$$

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

$$\begin{pmatrix} H_{12}(f)S_2(f,m) \\ H_{13}(f)S_3(f,m) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \frac{H_{22}}{H_{12}} & \frac{H_{23}}{H_{13}} \end{pmatrix}^{-1} \begin{pmatrix} X_1(f,m) \\ X_2(f,m) \end{pmatrix} \tag{10}$$

That is, the separated signals $Z_i(f,m)$ are obtained as follows:

$$\begin{bmatrix} Z_2(f,m) \\ Z_3(f,m) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \frac{H_{22}}{H_{12}} & \frac{H_{23}}{H_{13}} \end{bmatrix}^{-1} \begin{bmatrix} X_1(f,m) \\ X_2(f,m) \end{bmatrix} \tag{11}$$

Moreover, in this area, because the signals $H_{12}(f)$ $S_2(f,m)$ and $H_{13}(f)S_3(f,m)$ are not too greatly zero-padded due to the wide mask, 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 very large.

We proceed in the same way when $S_3(f,m)$ is zero and obtain the estimates of $H_{11}(f)S_1(f,m)$ and $H_{12}(f)S_2(f,m)$. It should be noted that, in Fig. 3, we have already confirmed that we do not often have three sources active simultaneously. We could also extend our method to a general case where there are N sources and M sensors. In such a case, we would write (6) as:

$$\begin{pmatrix} X_1(f,m) \\ \vdots \\ X_M(f,m) \end{pmatrix} = \mathbf{H'}(f) \begin{pmatrix} H_{k1}(f)S_1(f,m) \\ \vdots \\ H_{kN}(f)S_N(f,m) \end{pmatrix} \tag{12}$$

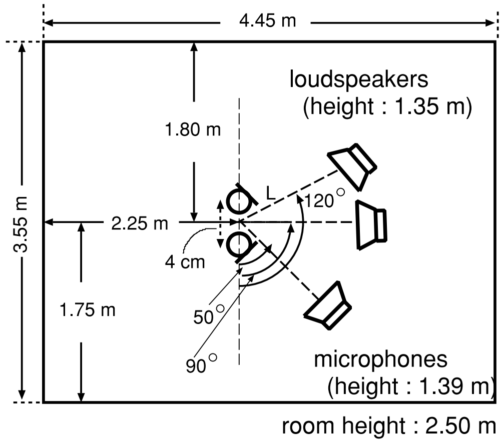


Fig. 7 Experimental conditions.

where

$$\mathbf{H'}_{ji}(f) = \begin{cases} 1 & j = k, \\ \frac{H_{ji}}{H_{ki}} & j \neq k. \end{cases} \tag{13}$$

and $\frac{H_{ji}}{H_{ki}}$ could be estimated by $\mathbf{E}[X_j(f)/X_k(f)]$ for an area i. Then we define a wide area so that we can re-define the square part of $\mathbf{H'}(f)$ as in (9), then separate the signals in the wide area by inverting the square part of $\mathbf{H'}(f)$ as we did in (10).

5. Experiments

5.1 Experimental Conditions

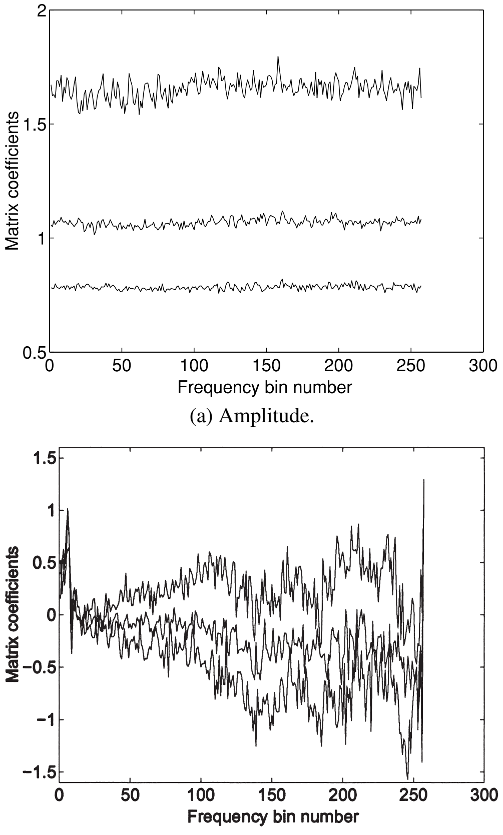
The recordings were made in a room with reverberation (TR=130 ms) using a two-element array of directional microphones 4 cm apart. The speech signals, sampled at 8 kHz, came from three directions: 120° (male), 90° (male) and 50° (female) and the distance between the sources and the sensors was L = 55 cm (see, Fig 7).

5.2 Stability of Estimated Mixing Matrix Coefficients

To evaluate the efficiency of our method, we need to know about the stability of the mixing matrix we estimated in the 2nd stage. In Fig. 8, we plot the amplitude and phase of the three coefficients $H_{2i}(f)/H_{1i}(f)$ $(i = 1, 2, 3)$ in (8). It is seen that our estimation generally offers a great stability in the whole, except for the low frequencies, where the time delay between the two microphones, which are positioned very close to each other, is harder to calculate with accuracy. However we can observe the constant amplitude and linear phase of the coefficients.

5.3 Mask Justification

Figure 9 justifies our decision to use wide masks. Indeed if we use narrow masks (e.g., area 3 in Fig. 6) as in the previous method, the recoverable power is only around 60% with



(a) Amplitude.

(b) Phase.

Fig. 8 Representation of the matrix coefficients, male $H_{23}(f)/H_{13}(f)$ - male $H_{22}(f)/H_{12}(f)$ - female $H_{21}(f)/H_{11}(f)$ combination, DFT size=512, TR=130 ms.

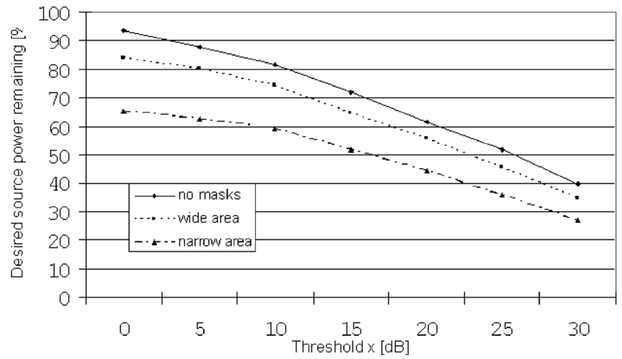


Fig. 9 Approximate WDO against the threshold, DFTsize=512, TR=130 ms.

a threshold of 10 dB whereas if we utilize wider masks (e.g., area 23 in Fig. 6), we can recover over 75% of this power. Consequently the technique using wide areas makes it possible to reduce the distortion of the separated signals, which was our aim.

Table 1 SIR and SDR calculated in dB for different approaches, DFT-size=512, TR=130 ms.“sparseness” means that performances are evaluated using the narrow masks; for “area12” the wide masks are used but not the mixing matrix (e.g., in the area comprising speech signal 1 and 2); for “H^{-1}_{area12}” our mixing matrix is used over area 12. Likewise “area23” and “H^{-1}_{area23}” are defined in area 23 comprising speech signals 2 and 3.

	SIR$_1$	SIR$_2$	SIR$_3$	SDR$_1$	SDR$_2$	SDR$_3$
sparseness	15.3	9.9	10.6	8.4	10.3	3.4
area12	0.5	2.3		10.5	12.0	
area23		4.2	−2.7		13.1	6.5
H^{-1}_{area12}	11.6	3.1		**8.7**	12.2	
H^{-1}_{area23}		3.3	7.6		12.5	**7.2**

Table 2 SIR and SDR calculated in dB for different approaches, DFT-size=512, TR=200 ms. “sparseness” means that performances are evaluated using the narrow masks; for “area12” the wide masks are used but not the mixing matrix (e.g., in the area comprising speech signal 1 and 2); for “H^{-1}_{area12}” our mixing matrix is used over area 12. Likewise “area23” and “H^{-1}_{area23}” are defined in area 23 comprising speech signals 2 and 3.

	SIR$_1$	SIR$_2$	SIR$_3$	SDR$_1$	SDR$_2$	SDR$_3$
sparseness	8.6	5.7	11.7	0.9	3.4	0.7
area12	0.6	0.5		5.4	5.9	
area23		−0.3	1.9		7.6	7.1
H^{-1}_{area12}	4.7	−1.1		**2.4**	5.5	
H^{-1}_{area23}		0.4	9.5		7.6	**4.6**

5.4 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:

$$SIR_i = 10\log\frac{\sum_n z_{is_i}(n)^2}{\sum_{i\neq j}\sum_n z_{is_j}(n)^2} \tag{14}$$

$$SDR_i = 10\log\frac{\sum_n x_{ks_i}(n)^2}{\sum_n (x_{ks_i}(n) - \alpha z_{is_i}(n-\phi))^2} \tag{15}$$

where the permutation is solved before calculating SIR and SDR, i.e. $z_i(t)$ is the estimation of $s_i(t)$. z_{is_j} is the output of the whole separating system at z_i when only s_j is active and s_k ($k \neq j$) does not exist, that is, $z_{is_j}(n) = u_{ij}(n) * s_j(n)$ where * denotes the convolution product, $u_{ij}(n)$ the impulse response from source s_j to separated signal z_i ($u_{ij}(n) = \sum_{l=1}^{M} w_{il}(n) * h_{lj}(n)$ with w_{il} the impulse response from a sensor l to a separated signal i). And x_{ks_j} is the observation obtained by microphone k when only s_j exists, i.e., $x_{ks_j}(n) = h_{kj}(n) * s_j(n)$. α is a constant that compensates for the amplitude difference and ϕ is an angle that fits the phase difference between input x_{ks_i} and output z_{is_i}. To evaluate the previous method (sparseness only method), we calculated SIR and SDR using both microphones’ measurements, and adopted the better values.

5.5 Separation Results

Tables 1 and 2 shows the results we obtained from our measurements for different degrees of reverberation. By “sparseness” we imply that we are evaluating the performances of our speech signals when employing the narrow masks. By “area12,” we mean that the performances are evaluated for the wide masks, e.g., in the area comprising speech signal 1 and 2 without using our mixing matrix. Finally, by "H^{-1}_{area12}," we mean that we apply our mixing matrix to area 12. Likewise "area23" and "H^{-1}_{area23}" are defined in area 23 comprising speech signals 2 and 3. Actually, we are comparing the conventional method with our proposed method in the different areas exploited.

As can be seen, the use of our proposed method allows us to obtain less distorted signals. In areas 12 and 23, the SDR values are high, however, the SIR values in these areas are bad as they consist of the mixtures two signals, which are not separated. We performed informal listening tests and it is important to note that, for the first signal in Table 1, although the two SDR results are very close, much less musical noise is heard when separation is undertaken using our approach rather than when only sparseness is used. To illustrate this point, we propose some sound samples on our web site [18].

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 experimental results are very encouraging in terms of quality and suggest that the combination of a time-frequency mask and a mixing matrix estimation is an approach that deserves serious investigation.

References

[1] S. Haykin ed., Unsupervised Adaptive Filtering, John Wiley & Sons, 2000.

[2] T.W. Lee, Independent Component Analysis—Theory and Applications, Kluwer, 1998.

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

[4] S. Amari, S.C. Douglas, A. Cichocki, and H.H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications, pp.101–104, April 1997.

[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, vol.22, pp.21–34, Nov. 1998.

[6] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," IEEE Trans. Speech Audio Process., vol.8, no.3, pp.320–327, May 2000.

[7] 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.

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

[9] 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, pp.3049–3052, 2002.

[10] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process., vol.52, no.7, pp.1830–1847, 2004.

[11] M. Zibulesky, B.A. Pearlmutter, P. Bofill, and P. Kisilev, "Blind source separation by sparse decomposition in a signal dictionary," TR no.99-1, FEC 313, University of New Mexico, Albuquerque, July 1999.

[12] F.J. Theis, C.G. Puntonet, and E.W. Lang, "A histogram-based overcomplete ICA algorithm," Proc. ICA2003, pp.1071–1076, 2003.

[13] F.J. Theis, E.W. Lang, and C.G. Puntonet, "A geometric algorithm for overcomplete linear ICA," Neurocomputing, vol.56, pp.381–398, 2004.

[14] M. B-Zadeh, A. Mansour, C. Jutten, and F. Marvasti, "A geometric approach for separating several speech signals," Proc. ICA2004, Lecture Notes in Computer Science 3195, pp.798–806, Springer-Verlag, 2004.

[15] A. Blin, S. Araki, and S. Makino, "Blind source separation when speech signals outnumber sensors using a sparseness-mixing matrix estimation," Proc. IWAENC2003, pp.211–214, 2003.

[16] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Blind separation of more speech than sensors with less distortion by combining sparseness and ICA," Proc. IWAENC2003, pp.271–274, 2003.

[17] 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.

[18] http://www.kecl.ntt.co.jp/icl/signal/araki/B3M.html

Audrey Blin received the Diplôme d'Ingénieur in electrical engineering in 2003 from the Ecole Nationale Supérieure en Electricité Informatique et Radiocommunication de Bordeaux, France. In 2003, she did a 6 month-internship at NTT Communication Science Laboratories. Currently, she is with the INRS-EMT, University of Québec, Canada as a Master Student since Autumn 2003. Her research interests include blind source estimation, equalization and separation of numerics and audio signals.

Shoko Araki received the B.E. and the M.E. degrees in mathematical engineering and information physics from the University of Tokyo, Japan, in 1998 and 2000, respectively. In 2000, she joined NTT Communication Science Laboratories, Kyoto. Her research interests include array signal processing, blind source separation applied to speech signals, and auditory scene analysis. She received the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, the Best Paper Award of the IWAENC in 2003 and the 19th Awaya Prize from Acoustical Society of Japan (ASJ) in 2001. She is a member of the IEEE and the ASJ.

Shoji Makino received the B.E., M.E., and Ph.D. degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981. He is now an Executive Manager at the NTT Communication Science Laboratories. He is also a Guest Professor at the Hokkaido University. His research interests include adaptive filtering technologies and realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech. He received the TELECOM System Technology Award of the TAF in 2004, the Best Paper Award of the IWAENC in 2003, the Paper Award of the IEICE in 2002, the Paper Award of the ASJ in 2005 and 2002, the Achievement Award of the IEICE in 1997, and the Outstanding Technological Development Award of the ASJ in 1995. He is the author or co-author of more than 200 articles in journals and conference proceedings and has been responsible for more than 150 patents. He is a member of the Conference Board of the IEEE SP Society and an Associate Editor of the IEEE Transactions on Speech and Audio Processing. He is also an Associate Editor of the EURASIP Journal on Applied Signal Processing. He is a member of the Technical Committee on Audio and Electroacoustics of the IEEE SP Society as well as the Technical Committee on Blind Signal Processing of the IEEE Circuits and Systems Society. He is also a member of the International ICA Steering Committee and the Organizing Chair of the ICA2003 in Nara. He is the General Chair of the IWAENC2003 in Kyoto. He was a Vice Chair of the Technical Committee on Engineering Acoustics of the IEICE and the ASJ. Dr. Makino is an IEEE Fellow, a council member of the ASJ, and a member of the EURASIP.