1—P—4

*Blind Separation of More Speech Signals than Sensors using Time-frequency Masking and Mixing Matrix Estimation

○Shoko Araki, △Audrey Blin , Shoji Makino (NTT Corporation)

1. Introduction

This paper focuses on underdetermined blind source separation (BSS) of three speech signals mixed in a real environment from measurements provided by two sensors. Up to now, solving the BSS problem in an underdetermined case has mainly consisted in assuming that the speech signals were sufficiently sparse [1-3]. They designed binary masks extracting signals at time-frequency points where only one signal was supposed to exist. However, due to unexpected discontinuous zero-padding, such separated signals have considerable distortion, and therefore a loud musical noise is heard. To overcome this issue, we propose to combine sparseness with a mixing matrix estimation. Experimental results in a real room show that our proposed method provides separated signals of better quality than those extracted with only the binary masks.

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 signals x_j $(j=1,...,M)$ can be modeled as convolutive mixtures of the source signals s_i $(i=1,..,N)$ as $x_j(t)=\sum_{i=1}^{N} h_{ji} * s_i(t)$ where h_{ji} is the impulse response from a source i to a sensor j. In this paper, we deal with the case of $N=3$ and $M=2$. 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 [1]. 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 ij 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 $\boldsymbol{S}(f,m)=[S_1(f,m), S_2(f,m), S_3(f,m)]^T$ are 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 consists in saying that the sources contain many zero samples. This 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.

4. Proposed method

In previous methods [1, 2], one of the major drawbacks was the occurrence of distortion, i.e., musical noise. To overcome this issue, we propose a three-step method.

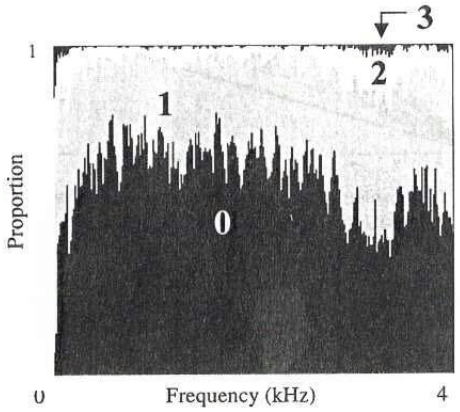

Fig. 1: Histogram of the number of active sources: 0, 1, 2 or 3 for a male-male-female combination for an fftsize of 512 in the reverberant case (T_R=200ms).

[1st step] One source detection

This first step consists in detecting the frame indices m when only one of the three sources is active for each frequency bin f. If the sources are sparse enough, we can find the time-frequency points where each source exists alone, they are located on each one of the three lines observed in the scatter-plot of the measurements (Fig. 2). By setting narrow areas surrounding these lines, such as areas 1, 2 and 3 in Fig. 2, we can determine when only one source is active. In the previous works [1-3], the signals are reconstructed using these time-frequency points. However, the separated signals had large distortion due to an unexpected zero-padding caused by the narrow masks.

[2nd step] Estimation of mixing matrix

Here, inspired by Deville's method [4], we estimate the mixing matrix in the time-frequency domain.

The observed signals can be rewritten as:

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

Using time-frequency points estimated in the first step when only S_i $(i=1, 2, 3)$ is active, we have $X_1(f,m)=H_{1i}(f)S_i(f,m)$ and $X_2(f,m)=H_{2i}(f)S_i(f,m)$. Then taking the ratio $X_2(f,m)/X_1(f,m)$ provides one of the factors of the mixing matrix $H_{2i}(f)/H_{1i}(f)$.

[3rd step] Separation of each two signals

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 because the mixing matrix is not square. Nevertheless, it is still possible to rebuild the time-frequency points when two sources are active by considering much wider areas like for example area 23 in Fig. 2. By taking such areas, we can estimate the (f, m)-points when $S_1(f, m)$ is null and our system becomes:

---

* 時間周波数マスキングと混合行列推定による 音源数 > マイク数の場合のブラインド音源分離, 荒木章子, A. Blin, 牧野昭二 (NTT 研究所)

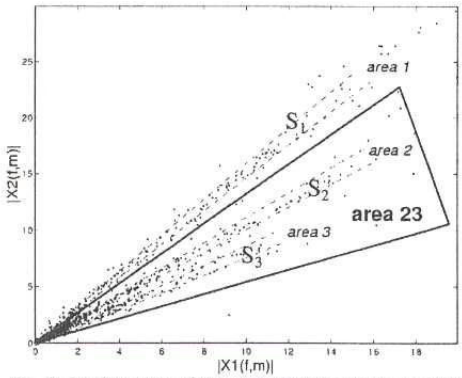

Fig. 2: Scatter-plots of the recorded signals for an fftsize of 512 and a frequency f of 312Hz. T_R=130 ms.

$$\begin{bmatrix} X_1(f,m) \\ X_2(f,m) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \frac{H_{22}(f)}{H_{12}(f)} & \frac{H_{23}(f)}{H_{13}(f)} \end{bmatrix} \cdot \begin{bmatrix} H_{12}(f)S_2(f,m) \\ H_{13}(f)S_3(f,m) \end{bmatrix} \quad (2).$$

Now the mixing matrix is square and can thus be inverted,

$$\begin{bmatrix} H_{12}(f)S_2(f,m) \\ H_{13}(f)S_3(f,m) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \frac{H_{22}(f)}{H_{12}(f)} & \frac{H_{23}(f)}{H_{13}(f)} \end{bmatrix}^{-1} \cdot \begin{bmatrix} X_1(f,m) \\ X_2(f,m) \end{bmatrix} \quad (3).$$

We proceed in the same way when $S_3(f,m)$ is null.

Note that, in Fig. 1, we have already confirmed that we do not often have three sources active simultaneously.

5. Experiments

5.1. Experimental conditions

The recordings were done in a room with little reverberation (T_R=130 ms) using a two-element array of directional microphones 4 cm apart. The speech signals came from three directions: 50° (female), 90° (male) and 120° (male) and the distance between the sources and the microphones was 55 cm. The sampling rate was 8 kHz and the FFT frame size was of 512, in which case the degree of speech signal overlap was the smallest [5].

To evaluate the separation performance, we calculated 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.

5.2. Mask justification

We assessed the approximate W-Disjoint Orthogonality [3] to check the percentage of recoverable power. The approximate WDO is defined as:

$$r_j(x) = 100. \| \phi_{(j,x)}(f,m) S_j(f,m) \|^2 / \| S_j(f,m) \|^2 \quad (4)$$

where

$$\phi_{(j,x)}(f,m) = \begin{cases} 1 & 20\log(|S_j(f,m)|/|Y_j(f,m)|) > x \\ 0 & otherwise \end{cases} \quad (5)$$

and where $Y_j(f,m)$ is the STFT of the summation of the sources interfering with source j: $y_j(t) = \sum_{i=1, i\neq j}^{N} s_i(t)$. If r_j is small, the energy of the output signal is lost by the binary mask. The solid line in Fig. 3 shows the approximate WDO when only the binary mask (5) is used. If we want a SIR of 20 dB, only around 50 % of the original power is recoverable, that is, almost half the points are zero-padded by the binary mask and such distortion cannot be avoided.

Figure 3 also shows the approximate WDO with narrow and wide masks. If we use narrow masks (e.g., area 3 in Fig. 2) as in the conventional method, the recoverable power

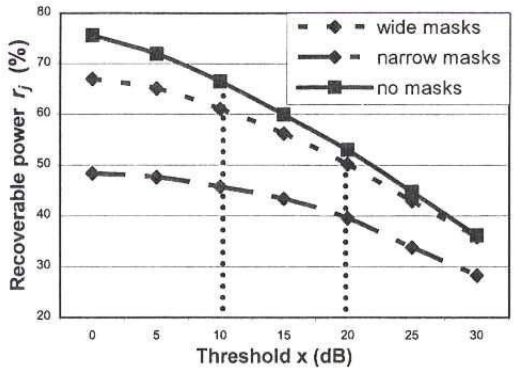

Fig. 3: Approximate WDO against the threshold x. fftsize=512, T_R=130 ms.

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
$invH_{12}$	11.6	3.1		**8.7**	12.2	
$invH_{23}$		3.3	7.6		12.5	**7.2**

is only around 45% with a threshold of 10dB, whereas if we utilize wider masks (e.g. area 23 in Fig. 2), we can recover over 60% of the power of original signals. Consequently the technique consisting in using wide areas makes it possible to reduce the distortion of the separated signals.

5.3. Results

Table 1 shows the separation results. "sparseness" implies the performance with conventional method, i.e., with the narrow masks (areas 1, 2, 3). "$invH_{12}$" and "$invH_{23}$" mean that we are applying our mixing matrix to area 12 and area 23, respectively. As we can see, the use of our method allows us to obtain less distorted signals without any serious deterioration of the SIR.

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 proposed method is an approach that deserves serious investigation.

References

[1] 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.

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

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

[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.