BLIND SOURCE SEPARATION BASED ON A BEAMFORMER ARRAY AND TIME FREQUENCY BINARY MASKING

[1,2]Jan Cermak, [1]Shoko Araki, [1]Hiroshi Sawada and [1]Shoji Makino

[1]NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
[2]Brno University of Technology, Faculty of Electrical Engineering and Communication,
Department of Telecommunication, Purkynova 118, 612 00 Brno, Czech Republic

ABSTRACT

This paper deals with a new technique for blind source separation (BSS) from convolutive mixtures. We present a three-stage separation system employing time-frequency binary masking, beamforming and a non-linear post processing technique. The experiments show that this system outperforms conventional time-frequency binary masking (TFBM) in both (over-)determined and underdetermined cases. Moreover it removes the musical noise and reduces interference in time-frequency slots extracted by TFBM.

Index Terms— Array signal processing, speech processing, speech enhancement

1. INTRODUCTION

Blind source separation has been intensively studied in recent years due to its many potential applications [1]. We address BSS as a technique used for separating signals from their mixtures when the mixing process is unknown. The separation techniques can basically be divided into two groups according to the number of sources N and sensors M, namely, (over-)determined cases when $N \le M$ and underdetermined cases when $N > M$.

Time-frequency binary masking [2] is a versatile approach capable of dealing with both (over-)determined and underdetermined cases that relies on signal sparseness and the assumption of independent sources. TFBM is based on estimating N time-frequency binary masks from the signal mixtures observed at M sensors. The target signals are then easily separated by multiplying the time-frequency binary masks with one of the signal mixtures. However TFBM leads to short tones randomly distributed in the separated target signal over the time and frequency and known as musical noise. Moreover interference remains in the extracted time-frequency slots.

We have proposed a system combining TFBM and adaptive beamformers [3] to ease the above-mentioned limitation. TFBM was exploited as a pre-separation process and the final separation was accomplished by multiple beamformers. Another approach to reducing musical noise is to employ a soft mask instead of a binary mask [4].

In this paper we present a technique that overcomes the limitations of TFBM described above. Our method *removes* the musical noise and suppresses the interference in *all* time-frequency slots. A block diagram of the proposed three-stage system is shown in Fig. 1. A pre-separation process using TFBM is employed to design the beamformers blindly as described in section 3. The core of the separation technique is the *beamformer array* (BA) described in section 4. In section 5 we present a method for signal enhancement, indicated as ENH in Fig. 1, that further improves the separation performance.

2. MIXING MODEL

The convolutive mixing model is used to describe signal observations in a real room environment, i.e. an environment with reverberations. We consider this model with M sensors observing N sources

$$x_j(t) = \sum_{k=1}^{\mathrm{N}} \sum_{l} h_{jk}(l) s_k(t-l), \qquad j = 1, \ldots, \mathrm{M}, \tag{1}$$

where $x_j(t)$ is the signal observed by the j-th sensor, t is the discrete time index, $s_k(t)$ is the k-th source signal and $h_{jk}(t)$ is the impulse response from the k-th source to the j-th sensor. The mixing model (1) in the time-frequency domain becomes

$$\mathbf{x}(f,\tau) \approx \sum_{k=1}^{\mathrm{N}} \mathbf{h}_k(f) s_k(f,\tau) \tag{2}$$

where $\mathbf{x}(f,\tau) = [x_1(f,\tau), \ldots, x_{\mathrm{M}}(f,\tau)]^{\mathrm{T}}$ is an observation vector and $\mathbf{h}_k(f) = [h_{1k}(f), \ldots, h_{\mathrm{M}k}(f)]^{\mathrm{T}}$ is a mixing vector. If the source signals are sparse, which holds for speech signals [5, 2], the sources rarely overlap in the time-frequency domain and (2) can be approximated as

$$\mathbf{x}(f,\tau) \approx \mathbf{h}_k(f) s_k(f,\tau), \tag{3}$$

where $s_k(f,\tau)$ is the dominant source at the time-frequency point (f,τ).

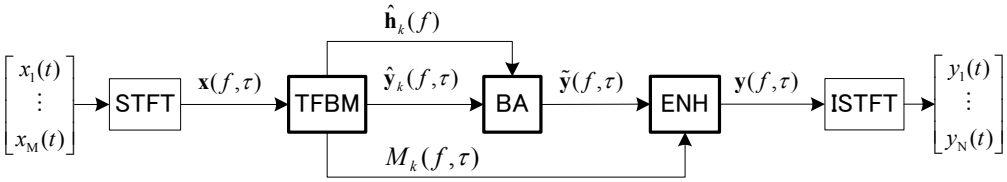

Fig. 1. Block diagram of the proposed system. STFT = short time Fourier transform, ISTFT = inverse STFT.

3. BEAMFORMER DESIGN WITHOUT A PRIORI INFORMATION

The main part of our separation system is the BA, which consists of minimum variance (Frost) beamformers. This section describes how to design the Frost beamformers blindly. The spatial filter for enhancing source $s_k(t)$ is designed by using [6],

$$\mathbf{w}_k(f) = \left(\frac{\mathbf{R}^{-1}(f)\mathbf{a}_k(f)}{\mathbf{a}_k{}^{\mathrm{H}}(f)\mathbf{R}^{-1}(f)\mathbf{a}_k(f)} \right), \tag{4}$$

where $\mathbf{R}(f) = \mathrm{E}[\mathbf{x}(f,\tau)\mathbf{x}^{\mathrm{H}}(f,\tau)]$ is the correlation matrix of the observation vector, E[] is the expectation operator, $^{\mathrm{H}}$ is the complex conjugate transpose and $\mathbf{a}_k(f) = [\mathrm{e}^{-\mathrm{j}2\pi f\tau_{1k}},...,\mathrm{e}^{-\mathrm{j}2\pi f\tau_{\mathrm{M}k}}]$ is the given steering vector. τ_{jk} is the time delay between the arrival of $s_k(t)$ at sensor j and at reference sensor J. The critical problem with the Frost beamformer (4) is that the design is not performed blindly.

To design the beamformer blindly and to further improve its performance, we have proposed using TFBM to estimate the mixing vector $\hat{\mathbf{h}}_k(f)$ and the correlation matrix of the observation vector in the jammer only period $\hat{\mathbf{R}}_k(f)$ [3]. Jammer only period is a time period when the target signal $s_k(t)$ is not active. Note that $\mathbf{R}(f)$ and $\mathbf{a}_k(f)$ can be substituted by $\hat{\mathbf{R}}_k(f)$ and $\hat{\mathbf{h}}_k(f)$ in (4), respectively. The remainder of this section reviews the way to estimate $\hat{\mathbf{R}}_k(f)$ and $\hat{\mathbf{h}}_k(f)$.

First, we estimate time-frequency binary mask $M_k(f,\tau)$ so that the pre-separated signal

$$\hat{\mathbf{y}}_k(f,\tau) = [\hat{y}_{1k}(f,\tau),...,\hat{y}_{\mathrm{M}k}(f,\tau)]^{\mathrm{T}} = M_k(f,\tau)\mathbf{x}(f,\tau) \tag{5}$$

becomes an estimation of the source, i.e. $\hat{\mathbf{y}}_k(f,\tau) \approx \mathbf{h}_k(f,\tau)s_k(f,\tau)$. $M_k(f,\tau)$ extracts the time-frequency slots of cluster C_k whose members are estimated to belong to the source signal $s_k(f,\tau)$

$$M_k(f,\tau) = \begin{cases} 1 & \mathbf{x}(f,\tau) \in \mathrm{C}_k \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

The cluster C_k can be estimated by using, for example, DUET [2], clustering the normalized observation vector [7] or the direction-of-arrival estimations [8]. Note that unlike conventional TFBM [2, 7] where the binary mask is applied to one observation only, the source separation (5) is accomplished over the whole observation vector $\mathbf{x}(f,\tau)$ in order to use the output vector $\hat{\mathbf{y}}_k(f,\tau)$ for designing the Frost beamformers.

Then we estimate $\hat{\mathbf{R}}_k(f)$ and $\hat{\mathbf{h}}_k(f)$ for the beamformer (4). Let $\hat{\mathbf{y}}_k(f,\tau)$ be the target component, and $\hat{\mathbf{y}}_b(f,\tau), b = (1,\ldots,\mathrm{N}, b \neq k)$ be interference for the source k. The correlation matrix $\hat{\mathbf{R}}_k(f)$ is estimated by

$$\hat{\mathbf{R}}_k(f) = \mathrm{E}[\hat{\mathbf{n}}_k(f,\tau)\hat{\mathbf{n}}_k{}^{\mathrm{H}}(f,\tau)] \tag{7}$$

where $\hat{\mathbf{n}}_k = \sum_{b \in z_k} \hat{\mathbf{y}}_b(f,\tau)$, and z_k is a set of jammer indexes whose size is less than or equal to M-1 [3]. For example if N=M=3 and k=1 then $z_k = \{2, 3\}$.

To estimate the mixing vector $\hat{\mathbf{h}}_k(f)$, we want to minimize the criterion

$$\Im(\mathbf{h}_k(f)) = \mathrm{E}[\mathbf{x}(f,\tau) - \hat{\mathbf{h}}_k(f)\hat{y}_{\mathrm{J}k}(f,\tau)]. \tag{8}$$

Setting the derivative $\partial\Im(\mathbf{h}_k(f))/\partial\mathbf{h}_k(f)$ to zero leads to the solution

$$\hat{\mathbf{h}}_k(f) = \frac{\mathrm{E}[\mathbf{x}(f,\tau)\hat{y}^*_{\mathrm{J}k}(f,\tau)]}{\mathrm{E}[|\hat{y}_{\mathrm{J}k}(f,\tau)|^2]}, \tag{9}$$

where * denotes conjugation.

4. BEAMFORMER ARRAY

In [3] and Section 3, we utilized just one beamformer $\mathbf{w}_k(f)$ to obtain the k-th output $y_k(f,\tau)$, which corresponds to an estimation of the target $s_k(f,\tau)$. In this paper, we propose BA, see Fig. 2, which utilizes D beamformers (k-th column in BA in Fig. 2) to estimate the k-th target signal $y_k(f,\tau)$.

Let us assume that our goal is to extract just one signal from the signal mixture. In this case BA consists only of one column (D beamformers). The main idea behind BA is to compose D different mixtures from the pre-separated signals provided by TFBM, which are then filtered by D beamformers. All these beamformers are designed to enhance the target signal. Each input mixture includes the pre-separated target signal and different pre-separated jammers. The important point is that all the jammers must be used at least once. As a result of beamforming we obtain the enhanced target signal D times. Finally, we add together all the outputs of the beamformers. Note that we used all pre-separated signals (all time-frequency slots) and therefore the BA output contains no musical noise.

Now, we describe the BA mathematically. The BA has N columns and D rows. D is a binomial coefficient estimated by

$$\mathrm{D} = \binom{\mathrm{N}-1}{\mathrm{Z}} = \frac{(\mathrm{N}-1)!}{\mathrm{Z}!(\mathrm{N}-1-\mathrm{Z})!}. \tag{10}$$

Each cell of the BA is a beamformer given as

$$\mathbf{w}_{kd}(f) = \left(\frac{\hat{\mathbf{R}}_{kd}^{-1}(f)\hat{\mathbf{h}}_k(f)}{\hat{\mathbf{h}}_k^{\mathrm{H}}(f)\hat{\mathbf{R}}_{kd}^{-1}(f)\hat{\mathbf{h}}_k(f)} \right), \qquad (11)$$

where d is the row index and k is the column index, and the index of the target signal i.e. each column is responsible for separating one target signal.

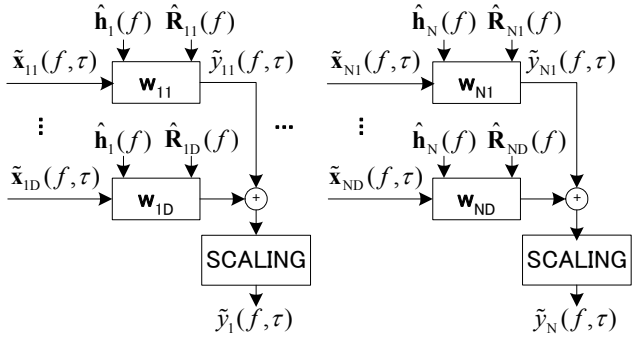

Fig. 2. Beamformer Array.

The input of a single beamformer is

$$\tilde{\mathbf{x}}_{kd}(f,\tau) = \hat{\mathbf{y}}_k(f,\tau) + \hat{\mathbf{n}}_{kd}(f,\tau), \qquad (12)$$

where $\hat{\mathbf{n}}_{kd}(f,\tau)$ is a vector of a mixture of $Z \in \{1,\ldots,M-1\}$ jammer signals. $\hat{\mathbf{n}}_{kd}(f,\tau)$ is counted by

$$\hat{\mathbf{n}}_{kd}(f,\tau) = \sum_{b \in z_{kd}} \hat{\mathbf{y}}_b(f,\tau), \qquad (13)$$

where z_{kd} is the d-th Z-subset [9] of θ_k. The set $\theta_k = \{q | 1\ldots N, q \neq k\}$ contains indexes of all the jammer signals for the k-th target signal. $\hat{\mathbf{R}}_{kd}(f) = \mathrm{E}[\hat{\mathbf{n}}_{kd}(f,\tau)\hat{\mathbf{n}}_{kd}^{\mathrm{H}}(f,\tau)]$ is estimated similarly to (7) from a mixture of jammers.

Let us take an example where N=4, M=3, Z=1 and k=1. The size of the BA is N=4 and D=3. The set of all jammer indexes is $\theta_1 = \{2, 3, 4\}$ and the 1-subsets are z_{11}=2, z_{12}=3, z_{13}=4. The input signals of the beamformers are $\tilde{\mathbf{x}}_{11}(f,\tau) = \hat{\mathbf{y}}_1(f,\tau) + \hat{\mathbf{y}}_2(f,\tau)$, $\tilde{\mathbf{x}}_{12}(f,\tau) = \hat{\mathbf{y}}_1(f,\tau) + \hat{\mathbf{y}}_3(f,\tau)$ and $\tilde{\mathbf{x}}_{13}(f,\tau) = \hat{\mathbf{y}}_1(f,\tau) + \hat{\mathbf{y}}_4(f,\tau)$.

In one column, the output signals $\tilde{y}_{kd}(f,\tau) = \mathbf{w}_{kd}^{\mathrm{H}}(f)\tilde{\mathbf{x}}_{kd}(f,\tau)$ of each beamformer are added together and then scaled with the number of rows D using block SCALING in Fig. 2

$$\tilde{y}_k(f,\tau) = \frac{\sum_{d=1}^{D} \tilde{y}_{kd}(f,\tau)}{D}, \qquad (14)$$

where $\tilde{\mathbf{y}}(f,\tau) = [\tilde{y}_1(f,\tau),\ldots,\tilde{y}_N(f,\tau)]$ is the output vector of the BA. Single beamformer outputs can be added due to the constraint of the Frost beamformer $\mathbf{w}_{kd}^{\mathrm{H}}(f)\mathbf{h}_k(f) = 1, \forall d$. The summation (14) is important for two reasons. Firstly, the target signal is enhanced since, unlike the jammers, it is in phase. Secondly, the musical noise is removed because all N pre-separated signals $\hat{\mathbf{y}}_k(f,\tau)$ are included in the input signals $\tilde{\mathbf{x}}_{k1}(f,\tau),\ldots,\tilde{\mathbf{x}}_{kD}(f,\tau)$ and therefore $\tilde{y}_k(f,\tau)$ includes no zero padding.

5. ENHANCEMENT

In the third section, ENH, the output vector $\tilde{\mathbf{y}}(f,\tau)$ of BA, is further enhanced. We propose a one-channel enhancement nonlinear technique based on knowledge of the time-frequency binary mask $M_k(f,\tau)$. The enhancement improves the interference suppression in the time-frequency slots of the target signal $\tilde{y}_k(f,\tau)$ where $M_k(f,\tau) = 0$. This can be easily achieved by

$$y_k(f,\tau) = \xi_k(f,\tau)\tilde{y}_k(f,\tau), \qquad (15)$$

where the enhancement factor $\xi_k(f,\tau)$ is

$$\xi_k(f,\tau) = \begin{cases} \alpha & M_k(f,\tau) = 1 \\ \beta & M_k(f,\tau) = 0 \end{cases}. \qquad (16)$$

The coefficients α and β can assume three different combinations

1) $\alpha > 1, \beta = 1$,
2) $\alpha = 1, \beta < 1$,
3) $\alpha > 1, \beta < 1$.

There are more possible ways to determine these coefficients. Note that this stage may cause musical noise e.g. by setting $\beta = 0$.

6. EXPERIMENTS

We performed experiments for a determined case (DC) (M=3, N=3) and an underdetermined case (UDC) (M=3, N=4) in a room with a reverberation time of 120 ms, (see Fig. 3). The source signals were 5-second English and Japanese utterances. The STFT frame size was L=512, the frame shift was L/4, and the sampling frequency f_s was 8 kHz. The separation performance was evaluated in terms of the signal-to-interference ratio (SIR) and signal-to-distortion ratio (SDR)

$$\mathrm{SIR}_k = 10\log_{10} \frac{\mathrm{E}\left[y_{kk}(t)^2 \right]}{\mathrm{E}\left[\sum_{b=1, b\neq k}^{N} y_{kb}(t)^2 \right]} [\mathrm{dB}], \qquad (17)$$

$$\mathrm{SDR}_k = 10\log_{10} \frac{\mathrm{E}\left[x_{Pk}(t)^2 \right]}{\mathrm{E}\left[\left(x_{Pk}(t) - \gamma y_{kk}(t-\Delta) \right)^2 \right]} [\mathrm{dB}], \qquad (18)$$

where $y_{kb}(t)$ are the jammer components that appear in the k-th output target signal, $y_{kk}(t)$ is the output signal without any contribution from the jammers and $x_{kP}(t) = \sum_l h_{Pk}(l)s_k(t-l)$. P is an arbitrarily selected sensor. Coefficients γ and Δ compensate the amplitude and delay, respectively.

The average results for the DC, when sources S_1, S_2 and S_3 were used, and for the UDC are summarized in the Table 1. We used TFBM as a performance reference for BA evaluation. The size of the BA was N=3, D=3 in DC and N=4, D=3 in the UDC. We use the notation BAZ-β in Table 1, where $Z \in \{1,2\}$ is the size of Z-subsets and β is a constant

enhancement value (16) for all time-frequency slots. α was set at one i.e. we used the second possible enhancement combination introduced in the previous section.

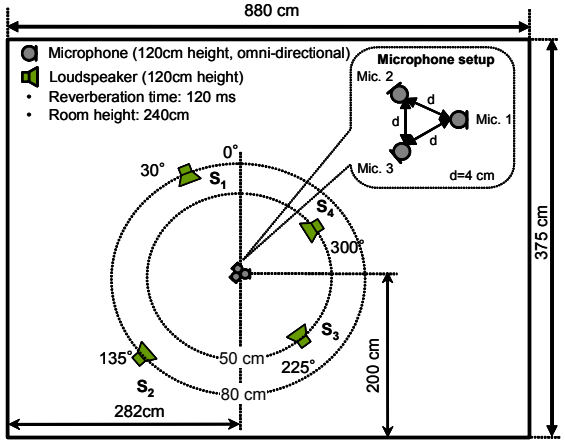

Fig. 3. Room setup.

We can see from Table 1 that the BA without enhancement (BA1-1 and BA2-1) outperforms conventional TFBM [7] in both the DC and UDC. BA1-1 achieves a higher SIR than BA2-1 because there are fewer jammers in $\tilde{\mathbf{x}}_{kd}(f,\tau)$ but on the other hand BA1-1 causes more distortion. The musical noise is removed in both setups.

Applying the third stage ENH of our system, we can see that the SIR increases further with decreasing β. But we can also observe the decreasing SDR and furthermore the musical noise becomes audible. By setting β=0 the BA output signals are binary masked, which is similar to the output with conventional TFBM, i.e. we introduce a musical noise into the BA output signals. However, unlike TFBM, our system reduces the interference even in time-frequency slots, where $M_k(f,\tau)=1$. This setup results in a lower SDR but a much higher SIR than with conventional TFBM. It is obvious from Table 1 that BA2-0 outperforms BA1-0, which means that BA2-0 has less interference in $M_k(f,\tau)=1$ time-frequency slots.

Table 1. Separation results.

	DC		UDC	
	SIR [dB]	SDR [dB]	SIR [dB]	SDR [dB]
TFBM	10.93	10.57	9.56	8.97
BA1-1	16.67	10.93	13.81	9.48
BA1-0.5	18.39	9.93	14.93	8.60
BA1-0	19.43	8.89	15.57	7.74
BA2-1	14.61	13.62	11.00	11.44
BA2-0.5	18.80	12.04	13.93	9.75
BA2-0	23.77	9.38	16.44	7.89

We can also see from Table 1 that the BA1-β setup is much less influenced by β changes than the BA2-β setup. Furthermore by decreasing β, the musical noise becomes audible sooner in the BA1-β setup. We can say that the BA with Z=1 better suppresses the jammers in $M_k(f,\tau)=0$ than the BA with Z=M-1. In contrast, the BA with Z=M-1 suppresses the jammers better in $M_k(f,\tau)=1$.

These results were anticipated and we can draw the general conclusion that a BA designed with Z=1 gives satisfactory results even as a two-stage system without enhancement. The BA designed with $Z=M-1$ has more interference in the time-frequency slots where $M_k(f,\tau)=0$ hence the third stage highly improves the performance.

7. CONCLUSION

We introduced a new three-stage BSS approach that provides high separation performance. We have shown that a BA removes the musical noise caused by conventional TFBM. Moreover the interference in the extracted time-frequency slots of the target signal is reduced. The third stage of our system allows us to control the level of musical noise and interference in the output signal. This is a versatile approach to signal enhancement that is capable of further improvement.

8. REFERENCES

[1] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*, John Wiley & Sons, New York, 2000.

[2] O. Yilmaz, S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on SP*, vol. 52, no. 7, pp. 1830-1847, 2004.

[3] J. Cermak, S. Araki, H. Sawada and S. Makino, "Blind speech separation by combining beamformers and time frequency binary mask," in *Proc. IWAENC2006*, Sept. 2006.

[4] S. Araki, H. Sawada, R Mukai and S. Makino, "Blind sparse source separation with spatially smoothed time-frequency masking," in *Proc. IWAENC2006*, Sept. 2006.

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

[6] D. Johnson, D. Dudgeon, *Array Signal Processing*, Prentice Hall, Upper Saddle River, 1993.

[7] S. Araki, H. Sawada, R. Mukai, S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. IWAENC2005*, pp. 117-120, Sept. 2005.

[8] Z. Yermeche, N. Grbic, I. Claesson, "Speech enhancement of multiple moving sources based on subband clustering time-delay estimation," in *Proc. IWAENC2006*, Sept. 2006.

[9] Mathworld – k-subset, Online: http://mathworld.wolfram.com /k-Subset.html, April 21, 2004.