BLIND SPEECH SEPARATION BY COMBINING BEAMFORMERS AND A TIME FREQUENCY BINARY MASK

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

[2]cermak4@kn.vutbr.cz
[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 describes a new method for blind speech separation (BSS) of convolutive mixtures. Our approach is based on a widely used speech enhancement method called beamforming. We utilize this technique for BSS by combining a beamformer and a time-frequency binary mask (TFBM) in one system. We propose two different approaches using the same basis but with a different setup. The first approach is designed for (over-)determined configurations, i.e. the number of sensors is equal to or greater than the number of sources. The second approach is designed for underdetermined configurations, i.e. the sources outnumber the sensors. Experimental results show that the proposed approach provides better results than the sole use of a conventional TFBM or a conventional beamformer.

1. INTRODUCTION

BSS aims to separate speech signals from their mixtures without any a priori information about the source position, room acoustics, mixing processes etc. Speech separation methods can be distinguished into two groups according to applicability for only an (over-)determined case, or for both (over-)determined and underdetermined cases.

A well known representative of (over-)determined methods is independent component analysis ICA [1,2]. ICA is a statistical BSS method relying only on statistical independence of the source signals. Another possible (over-)determined approach is a *multiple-beamformer*. Beamformer [3] performs spatial filtering by forming a directivity pattern of an array with M sensors in order to emphasize a target signal $s_k(t)$ arriving from a given direction and to suppress signals arriving from other directions (jammers). In order to separate N sources, a set of N different beamformers must be employed. However, this method may also be used for an underdetermined case but the jammer suppression is not efficient because only M-1 minimums (null patterns) can be designed in a directivity pattern.

The critical problem of set of multiple-beamformers or of beamformer itself is that the speech separation is *not blind*, because beamformer needs a priori information about the target signal $s_k(t)$, e.g. a mixing vector or at least its approximation (steering vector).

Time-frequency binary mask (TFBM) [4,5] is a BSS approach that can be applied even to an underdetermined case because the TFBM method relies on time-frequency sparseness. However it has a musical noise problem due to zero padding in the time-frequency domain.

In this paper, we present a BSS method that overcame the above mentioned limitations. Our method removes and significantly reduces musical noise for (over-)determined and underdetermined cases, respectively. This is achieved by combining multiple-beamformers and TFBM. We use TFBM for mixing vector estimation in order to make multiple-beamformers blind. Furthermore, TFBM is exploited for reducing the number of jammers so that only M-1 jammers are included in the signal mixtures at the beamformer input even in the underdetermined case. Beamformer then separates the target signal, thus the musical noise is reduced.

2. MIXING MODEL AND CONVENTIONAL APPROACHES

2.1. Mixing Model

We consider M sensors observing N sources as convolutive mixtures

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

where $x_j(t)$ is the signal observed by the j-th microphone, t is the discrete time index, $s_k(t)$ is the k-th source signal and $h_{jk}(t)$ represents 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 the vector of the impulse responses (mixing vector). If the source signals are sparse, which holds for speech signals [5], 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).

2.2. Conventional beamformer

A beamformer is a set of filters $w_{jk}(f)$ that perform spatial filtering. The enhanced output signal in the time-frequency domain is obtained as

$$y_k(f,\tau) = \mathbf{w}_k(f)^{\mathrm{T}} \mathbf{x}(f,\tau), \tag{4}$$

where $\mathbf{w}_k(f)=[w_{1k}(f),\ldots,w_{\mathrm{M}k}(f)]^{\mathrm{T}}$. We describe the minimum variance (Frost) beamformer [3]. The filters $w_{jk}(f)$ are designed by using

$$\mathbf{w}_k(f) = \left(\frac{\mathbf{R}_k(f)^{-1}\mathbf{a}_k(f)}{\mathbf{a}_k(f)^{\mathrm{H}}\mathbf{R}_k(f)^{-1}\mathbf{a}_k(f)} \right)^{*}, \tag{5}$$

where $^{\mathrm{H}}$ is the complex conjugate transpose, * is the conjugation, $\mathbf{R}_k(f)=\mathrm{E}[\mathbf{x}(f,\tau)\mathbf{x}(f,\tau)^{\mathrm{H}}]$ is the correlation matrix of observation vector $\mathbf{x}(f,\tau)$, and E[] is the mean operator. $\mathbf{a}_k(f)$ is a steering vector representing an approximation of $\mathbf{h}_k(f)$ for an anechoic environment

$$\mathbf{a}_k(f) = \left[\mathrm{e}^{-\mathrm{j}2\pi f \tau_{1k}}, \ldots, \mathrm{e}^{-\mathrm{j}2\pi f \tau_{\mathrm{M}k}} \right], \tag{6}$$

where τ_{jk} is the time delay between the arrival of $s_k(t)$ at sensor j and at reference sensor J. τ_{jk} is dependent on the target location, the sensor array geometry and the propagation velocity c. However, it is difficult to determine a precise steering vector due to the misalignment of the sensor array or reverberation in a real situation.

Instead of $\mathbf{a}_k(f)$, we can employ $\mathbf{h}_k(f)$, which contains the correct array geometry and reverberation information. Moreover, $\mathbf{R}_k(f)$ can be substituted by the correlation matrix of the observation vector in the jammer only period $\overline{\mathbf{R}}_k(f)$. These changes both result in improved performance. However, the $\mathbf{h}_k(f)$ measurement is not realistic in practice and furthermore the estimation of the jammer only period constitutes a difficult problem especially when the jammers are non-stationary signals.

2.3. Conventional TFBM

A conventional TFBM is shown in Fig. 1. The principle of the TFBM is introduced in [4, 5] and therefore we describe the TFBM briefly here. First, the time domain mixtures $x_j(t)$ are transformed into the time-frequency domain by the short-time Fourier transform (STFT). Observation vector $\mathbf{x}(f,\tau)$ is then normalized so that it forms clusters corresponding to individual sources $s_k(t)$. Normalization is undertaken in block NORMAL by selecting reference sensor J and counting

$$\breve{x}_j(f,\tau) = \left|x_j(f,\tau)\right| \exp\left[-\mathrm{j} \frac{\arg\left(x_j(f,\tau)/x_{\mathrm{J}}(f,\tau) \right)}{2\pi f c^{-1} d_{\max}} \right], \tag{7}$$

$$\overline{\mathbf{x}}(f,\tau) = \frac{\breve{\mathbf{x}}(f,\tau)}{\left\|\breve{\mathbf{x}}(f,\tau)\right\|}, \tag{8}$$

where $\mathrm{d}_{\max}$ is the maximal distance between the reference sensor J and sensor $j \in \{1,\ldots,\mathrm{M}\}$ and $\breve{\mathbf{x}}(f,\tau)=[\breve{x}_1(f,\tau),\ldots,\breve{x}_{\mathrm{M}}(f,\tau)]^{\mathrm{T}}$.

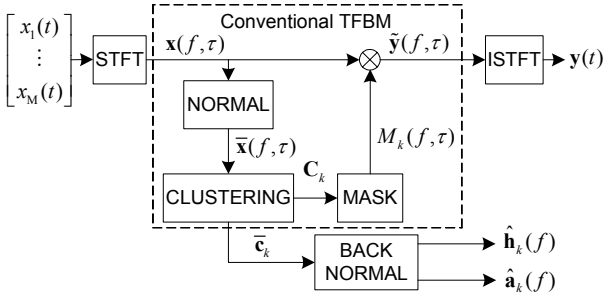

Figure 1. *Time-frequency binary mask (TFBM).*

Normalized vector $\overline{\mathbf{x}}(f,\tau)=[\overline{x}_1(f,\tau),\ldots,\overline{x}_M(f,\tau)]^{\mathrm{T}}$ is then clustered in order to find N clusters $\mathrm{C}_1,\ldots,\mathrm{C}_{\mathrm{N}}$ corresponding to individual source signals. Note that the number of sources N must be known beforehand. Clustering is achieved by minimizing the objective function

$$\Im = \sum_{k=1}^{\mathrm{N}} \sum\nolimits_{\overline{\mathbf{x}}(f,\tau)\in \mathrm{C}_k} d_k(f,\tau), \tag{9}$$

where $d_k(f,\tau)$ is the distance from the normalized vector $\overline{\mathbf{x}}(f,\tau)$ to the cluster centers $\mathbf{c}_k=[\,c_{1k,\ldots,}\,c_{\mathrm{M}k}]^{\mathrm{T}}$

$$d_k(f,\tau) = 1 - \mathrm{real}(\overline{\mathbf{x}}(f,\tau)^{\mathrm{T}} \mathbf{c}_k^{\,*}). \tag{10}$$

$\Im$ is minimized by the following updates

$$\mathrm{C}_k = \left\{ \overline{\mathbf{x}}(f,\tau) \middle| k = \underset{i}{\mathrm{argmin}}\,(d_i(f,\tau)) \right\}, \tag{11}$$

$$\overline{\mathbf{c}}_k = \mathrm{E}\left[\overline{\mathrm{x}}(f,\tau)\right]_{\overline{\mathbf{x}}\in \mathrm{C}_k}, \quad \mathbf{c}_k = \overline{\mathbf{c}}_k / \left\|\overline{\mathbf{c}}_k\right\|, \tag{12}$$

where $\mathrm{E}[]_{\overline{\mathbf{x}}\in \mathrm{C}_k}$ is the mean operator for the members of a cluster C_k. The vector of separated target signals $\tilde{\mathbf{y}}(f,\tau)=[\tilde{y}_1(f,\tau),\ldots,\tilde{y}_{\mathrm{N}}(f,\tau)]^{\mathrm{T}}$ is obtained by

$$\tilde{y}_k(f,\tau) = M_k(f,\tau) x_{\mathrm{P}}(f,\tau), \tag{13}$$

where $\mathrm{P} \in \{1,\ldots,\mathrm{M}\}$ is an arbitrary sensor index and $M_k(f,\tau)$ is a time-frequency binary mask extracting the time-frequency points of cluster C_k.

$$M_k(f,\tau) = \begin{cases} 1 & \overline{\mathbf{x}}(f,\tau) \in \mathrm{C}_k \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

Finally, the vector of the separated target signals

$\tilde{\mathbf{y}}(f,\tau)$ is transformed back into the time domain by inverse STFT (ISTFT).

3. PROPOSED APPROACH FOR (OVER-) DETERMINED CASE

Figure 2 shows the flow of our method. The time domain mixtures $x_j(t)$ are first transformed into the time-frequency domain by STFT. Observation vector $\mathbf{x}(f,\tau)$ is the beamformer and TFBM input. The function of the TFBM, the jammer selection (JS) block, and the correlation (CORR) block is to estimate the jammer correlation matrix $\hat{\mathbf{R}}_k(f)$ and $\hat{\mathbf{h}}_k(f)$ or $\hat{\mathbf{a}}_k(f)$, where ^ stands for *estimated* value. Finally, the vector of the separated target signals $\mathbf{y}(f,\tau)=[\,y_1(f,\tau),\ldots,y_N(f,\tau)]^T$ is transformed back into the time domain by ISTFT.

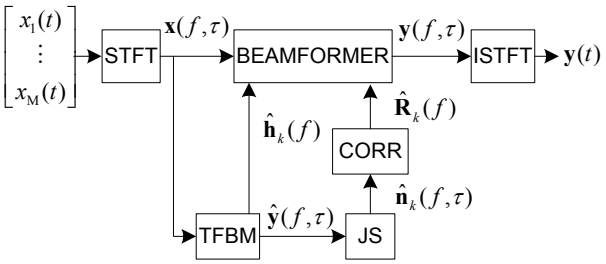

Figure 2. *Proposed method for (over-)determined case.*

Jammer correlation matrix $\hat{\mathbf{R}}_k(f)$ estimation:
In contrast to a conventional TFBM (13), we compute the separated signal matrix $\hat{\mathbf{y}}(f,\tau)=[\hat{\mathbf{y}}_1(f,\tau),\ldots,\hat{\mathbf{y}}_N(f,\tau)]$ where $\hat{\mathbf{y}}_k(f,\tau)=M_k(f,\tau)\mathbf{x}(f,\tau)^T=[\hat{y}_{1k}(f,\tau),\ldots,\hat{y}_{Mk}(f,\tau)]^T$. The separated signal matrix $\hat{\mathbf{y}}(f,\tau)$ is needed in order to satisfy the requirement of M channel signals for beamformer design. The jammer signals $\hat{\mathbf{n}}_k(f,\tau)=[\hat{n}_{1k}(f,\tau),\ldots,\hat{n}_{Mk}(f,\tau)]^T$ are estimated in the JS block as

$$\hat{n}_{jk}(f,\tau)=\sum_{b=1,b\neq k}^{N}\hat{y}_{jb}(f,\tau). \qquad (15)$$

Finally, $\hat{\mathbf{R}}_k(f)=\mathrm{E}(\hat{\mathbf{n}}_k(f,\tau)\hat{\mathbf{n}}_k(f,\tau)^H)$ is counted in the CORR block.

Mixing vector $\hat{\mathbf{h}}_k(f)$ estimation:
In our approach, we extend the TFBM to estimate the mixing vector $\hat{\mathbf{h}}_k(f)$ or steering vector $\hat{\mathbf{a}}_k(f)$. This is possible because the cluster centroids $\mathbf{c}_k$, obtained through the normalization and clustering process, represent an estimation of mixing vector $\mathbf{h}_k(f)$ [6]. This can be derived from (3), (7), (8) and (12) as

$$\bar{\mathbf{c}}_k=\mathrm{E}[\bar{\mathbf{x}}(f,\tau)]_{\mathbf{x}\in C_k}=\mathrm{E}[\bar{\mathbf{h}}_k(f)]_f, \qquad (16)$$

where $\bar{\mathbf{h}}_k(f)$ is a normalized mixing vector. As regards (16), the mixing vector can be obtained through back-normalization by using (8) and (7)

$$\hat{\mathbf{h}}_k(f)=\frac{|\bar{\mathbf{c}}_k|}{\sqrt{M}}\exp\left[-j2\pi f d_{max}c^{-1}\arg(\bar{\mathbf{c}}_k\bar{\mathbf{c}}_j)\right]. \qquad (17)$$

Steering vector $\hat{\mathbf{a}}_k(f)$ (6) can also be derived by substituting the time delay $\hat{\boldsymbol{\tau}}_k=d_{max}c^{-1}\arg(\bar{\mathbf{c}}_k\bar{\mathbf{c}}_j)$, obtained from (17), where $\hat{\boldsymbol{\tau}}_k=[\hat{\tau}_{1k},\ldots,\hat{\tau}_{Mk}]$.

4. PROPOSED APPROACH FOR UNDERDETERMINED CASE

The method proposed in the previous section may also be used for an underdetermined case but it results in low signal-to-interference ratio (SIR), see (19), as described in section 1. In order to achieve a high SIR, our underdetermined approach implements a *jammer reduction*, which reduces the number of jammers $\bar{N}$ in the signal mixtures at the beamformer input, so that $\bar{N}$=M-1. The flow of our proposed method is shown in Fig. 3.

Jammer reduction:
In Fig. 3, the JS block counts the *reduced jammer signal mixture vector* $\hat{\mathbf{n}}'_k(f,\tau)=[\hat{n}'_{1k}(f,\tau),\ldots,\hat{n}'_{Mk}(f,\tau)]$ and the *estimated* (pre-separated) *target signal* $\hat{\mathbf{y}}_k(f,\tau)$ that should be enhanced by the beamformer. The reduced jammer signal mixture $\hat{n}'_{jk}(f,\tau)$ is a summation of $\bar{N}$=M-1 jammer signals estimated by the TFBM. The jammer reduction is achieved by

$$\hat{n}'_{jk}(f,\tau)=\sum_{b=1,b\neq k,b\notin z_j}^{N}\hat{y}_{jb}(f,\tau), \qquad (18)$$

where $z_j\in\{1\ldots N\}^{N-M}$ is a set of N-M jammer indexes. Jammer indexes z_j represent jammer signals that should not be included in $\hat{n}'_{jk}(f,\tau)$. Jammer indexes are determined by the selection criterion. Different selection criteria can be used. In our approach we select the jammer signals with the fewest cluster members in order to minimize musical noise.
The beamformer input $\hat{\mathbf{x}}_k(f,\tau)=\hat{\mathbf{n}}'_k(f,\tau)+\hat{\mathbf{y}}_k(f,\tau)$ is the summation of the reduced jammer mixture $\hat{\mathbf{n}}'_k(f,\tau)$ and the estimated target signal $\hat{\mathbf{y}}_k(f,\tau)$.
Beamformer design is then based on the *reduced jammer correlation matrix* $\hat{\mathbf{R}}'_k(f)=\mathrm{E}[\hat{\mathbf{n}}'_k(f,\tau)\hat{\mathbf{n}}'_k(f,\tau)^H]$ and on the estimated mixing vector $\hat{\mathbf{h}}_k(f)$ or estimated steering vector $\hat{\mathbf{a}}_k(f)$.

5. EXPERIMENTS

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

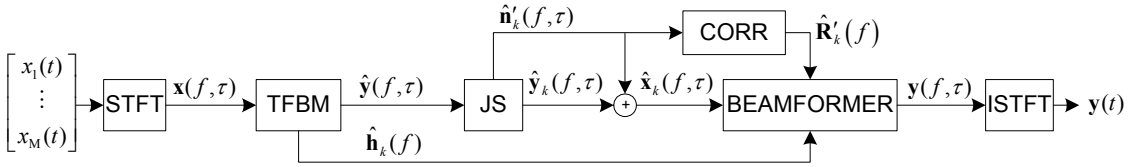

Figure 3. *Proposed method for underdetermined case.*

$$\mathrm{SIR}_k = 10\log_{10}\frac{\mathrm{E}\left[y_k(t)^2\right]}{\mathrm{E}\left[\sum_{i=1,i\neq k}^{\mathrm{N}} y_i(t)^2\right]}\,[\mathrm{dB}], \tag{19}$$

$$\mathrm{SDR}_k = 10\log_{10}\frac{\mathrm{E}\left[x_{\mathrm{P}k}(t)^2\right]}{\mathrm{E}\left[\left(x_{\mathrm{P}k}(t)-\alpha y_k(t-\mathrm{D})\right)^2\right]}\,[\mathrm{dB}], \tag{20}$$

where $y_i(t)$ are the jammer components that appear in the output target signal, $y_k(t)$ is the output signal without any contribution from the jammers and $x_{\mathrm{P}k}(t) = \sum_l h_{\mathrm{P}k}(l)s_k(t-l)$. Coefficients α and D compensate the amplitude and delay, respectively.

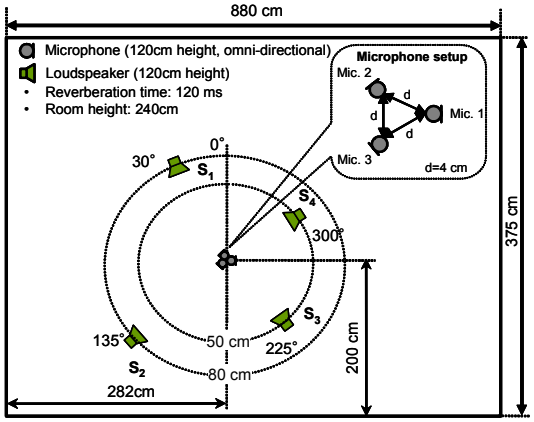

Figure 4. *Room setup.*

Table 1 shows the separation results for the determined case (D), when sources S_1, S_2 and S_3 were used, and for the underdetermined case (UD). We used different BSS approaches in order to compare them. A conventional beamformer (CB) was designed as described in section 2.2, namely using steering vector $\mathbf{a}_k(f)$ with given target locations and observation correlation matrix $\mathbf{R}_k(f)$. Note that separation is not performed blindly. The conventional TFBM setup corresponds to the approach outlined section 2.3. Finally, we used our proposed method, which exploits $\hat{\mathbf{h}}(f)$, $\hat{\mathbf{R}}_k(f)$ and $\hat{\mathbf{h}}(f)$, $\hat{\mathbf{R}}'_k(f)$ in the D and UD cases, respectively.

The CB setup did not achieve good results because the jammer only correlation matrix $\hat{\mathbf{R}}_k(f)$ was not used and furthermore the steering vector $\mathbf{a}_k(f)$ could not reflect the room reverberation or sensor misalignment. The TFBM achieved higher SIR and SDR values than the CB but the zero padding in the time-frequency domain means that we hear large musical noise, especially in the UD case. On the other hand, the proposed methods achieve higher separation performance with much less and without musical noise for UD and D cases, respectively. Furthermore, the limitations of the conventional approaches, described in section 1, are overcome.

Table 1. Experimental results for 3 microphone setting.

Design Method	Average SIR [dB]		Average SDR [dB]	
	D	UD	D	UD
CB	5.1	2.3	7.3	7.2
TFBM	10.9	9.6	10.6	9.0
Proposal	14.6	9.3	13.6	10.8

6. CONCLUSION

We introduced a new BSS approach for both (over-)determined and underdetermined cases by assuming source sparseness. Our method combines beamformers and TFBM and provides with better separation performance than conventional techniques. Because the beamformer is the core of the separation, the musical noise is significantly reduced and removed in underdetermined and (over-)determined cases, respectively. Furthermore, the computation time is not significantly increased revealing the potential for real time application.

7. REFERENCES

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

[2] H. Sawada, R. Mukai, S. Araki and S. Makino, "Frequency-domain blind source separation," in *Speech Enhancement*, J Benesty, S. Makino and J. Chen, Eds. Springer, Mar. 2005.

[3] D. Johnson, D. Dudgeon, *Array Signal Processing*, Prentice Hall, 1993.

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

[5] 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.

[6] S. Araki, H. Sawada, R. Mukai, S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. ICASSP2006*, May 2006.