Musical noise reduction in time-frequency-binary-masking-based blind source separation systems

[2, 3, a]J. Čermák, [1]S. Araki, [1]H. Sawada and [1]S. Makino

[1]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
[2]Institute of Radio Engineering and Electronics, Academy of Sciences, Prague, Czech Republic
[3]Department of Telecommunication, Brno University of Technology, Brno, Czech Republic
E-mail: cermak4@kn.vutbr.cz

Abstract: Blind source separation (BSS) problem consists of estimating N sources from M mixtures without using source and mixing information. In this paper we focus on improving BSS method called time-frequency binary masking (TFBM). TFBM is a versatile approach due to its ability to separate signals in both (over-)determined case ($N \leq M$) and underdetermined case ($N > M$). The cost we pay for the versatility is the musical noise which is introduced to the separated signals by the binary masking. We introduce a method, which reduces musical noise and improves separation performance in both (over-)determined and underdetermined cases.

1. Introduction

Blind source separation methods can be employed in many applications e.g. teleconferences, pre-processing for speech recognition. It has been discussed by many authors [1, 2] lately but most of the existing approaches are applicable just for a (over-)determined case (DC) [2]. Approaches for underdetermined case (UDC) suffer from musical noise [3] or can not be used in unknown environment. Our aim is to develop an approach that is blind, does not cause musical noise and can be used even for UDC; see table 1 for comparison between different source separation methods.

	DC	UDC	Musical noise	Blind
ICA	☺	☹	☺	☺
TFBM	☺	☺	☹	☺
AFB	☺	😐	☺	☹
Proposal	☺	☺	☺	☺

Table 1: Comparison of source separation methods.

Independent component analysis (ICA) [2, 4] is a statistical BSS method relying only on statistical independence of the source signals. ICA is a well known representative of DC which does not cause musical noise.

Time-frequency binary mask [3, 5] is a BSS approach that can be applied even to an UDC. However it causes a musical noise problem due to zero padding in the time-frequency (TF) domain.

Adaptive beamformer (ABF) [6] is a system performing 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 ABFs (multiple-beamformer) must be employed. The critical problem of ABF is that the speech separation is *not blind*,

[a] This work was done during the internship at NTT Communication Science Laboratories.

because ABF needs a priori information about the target signal $s_k(t)$, e.g. a mixing vector or at least its approximation (steering vector). Multiple-beamformer may also be used for an UDC but the jammer suppression is not efficient because only M-1 minimums (null patterns) can be designed in a directivity pattern.
In this paper we combine TFBM and multiple-beamformer into one system to reduce the musical noise caused by conventional TFBM while keeping its versatility. We employ the TFBM as a pre-separation part and the core of the separation is done by the multiple-beamformer.
We will focus on general description of the proposed system. The detailed description of single parts of our system [7] is beyond the scope of this article. The organization of this paper is following. First we give a brief overview about the mixing system in section 2. In section 3, the proposed system will be introduced and finally the experimental results will be shown in section 4.

2. Mixing process

We consider a mixing process that is taking place in unknown real environment i.e. we have to deal with convolutive mixtures

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

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.
By applying short time Fourier transform (STFT) the convolutive mixtures (1) are reduced to instantaneous mixtures at each frequency

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

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 an mixing vector.
In BSS problem the sparseness of the source signals is often considered. Sparse signals are signals which have only few samples different from zero. Speech signal in time-frequency (TF) domain can be considered as a sparse signal [3] because the energy in voiced segments is accumulated around the multiples of fundamental frequency resulting in low energy in all other TF slots. Further we consider that the sources $s_k(f,\tau)$ do not overlap

$$\prod_{k=1}^{\mathrm{N}} s_k(f,\tau) = 0, \ \ \forall f, \tau. \tag{3}$$

The assumption (3) is not satisfied for simultaneous speech signals but it can still be used as an approximation [3] for independent utterances because the speech signals are sparse in TF domain and each speaker has different speech characteristics. Using weakened version of (3) as an approximation the mixing model in TF domain (2) becomes

$$\mathbf{x}(f,\tau) \approx \mathbf{h}_k(f) s_k(f,\tau), \tag{4}$$

where $s_k(f,\tau)$ is the *dominant* source at the TF slot.

3. Proposed approach

We propose BSS method based on multiple-beamformer and TFBM. Block diagram of our proposed system, shown in Figure 1, extracts one of N signals from M mixtures. TFBM is exploited as a pre-separation process to design ABFs and to compose ABFs inputs. Each ABF extracts then one target signal $s_k(t)$.

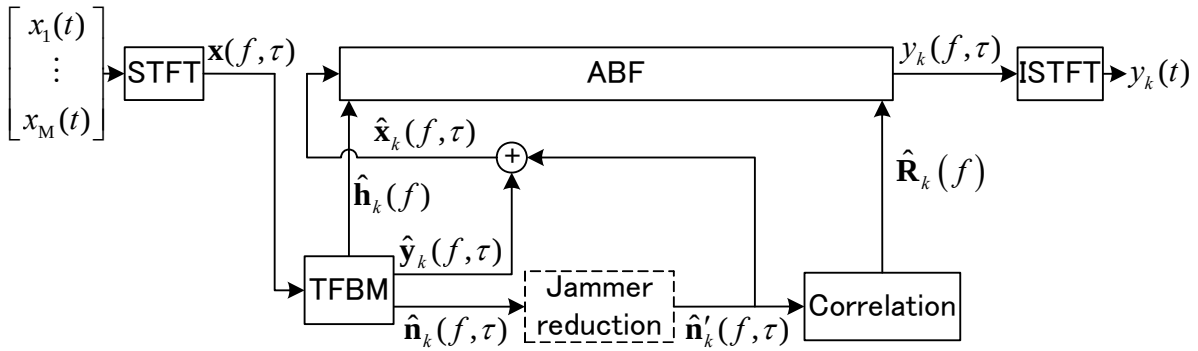

Figure 1: Block diagram of proposed BSS system; ISTFT = inverse STFT.

3.1. Adaptive beamformer

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

$$y_k(f,\tau) = \mathbf{w}_k(f)\hat{\mathbf{x}}_k(f,\tau), \tag{5}$$

where $\mathbf{w}_k(f) = [w_{1k}(f),\ldots,w_{Mk}(f)]$ and $\hat{\mathbf{x}}_k(f,\tau) = [\hat{x}_1(f,\tau),\ldots,\hat{x}_M(f,\tau)]^T$ is the input vector of ABF as described in section 3.3. Now let us consider that the mixing vector $\mathbf{h}_k(f)$ and the observation vector in the jammer only period (the time period when the source signal $s_k(t)$ is not active) is known. The filters $w_{jk}(f)$ can then be designed by using

$$\mathbf{w}_k(f) = \left(\frac{\mathbf{R}_k(f)^{-1}\mathbf{h}_k(f)}{\mathbf{h}_k(f)^H \mathbf{R}_k(f)^{-1}\mathbf{h}_k(f)} \right), \tag{6}$$

where ^{H} is the complex conjugate transpose, $\mathbf{R}_k(f) = \mathrm{E}[\mathbf{n}_k(f,\tau)\mathbf{n}_k(f,\tau)^H]$ is the correlation matrix of the observation vector in the jammer only period $\mathbf{n}_k(f,\tau)$ and E[] is the mean operator.

But mixing vector $\mathbf{h}_k(f)$ and correlation matrix of the observation vector in the jammer only period $\mathbf{R}_k(f)$ are usually not known because measurement of $\mathbf{h}_k(f)$ is not realistic in practice and the estimation of the jammer only period constitutes a difficult problem especially when the jammers are non-stationary signals. Therefore by conventional design of ABF [6] an approximation of mixing vector for an anechoic environment (steering vector) and correlation matrix of observation vector $\mathbf{R}(f) = \mathrm{E}[\mathbf{x}(f,\tau)\mathbf{x}(f,\tau)^H]$ is used instead of $\mathbf{h}_k(f)$ and $\mathbf{R}_k(f)$, respectively. This simplification is a standard design procedure although it lowers the separation performance and makes the *blind* design impossible.

Though all the difficulties we use (6) for designing ABF by estimating $\hat{\mathbf{h}}_k(f)$ and $\hat{\mathbf{R}}_k(f)$ in our proposed system; ^ stands for *estimated* value.

3.2. Multi-channel TFBM

Conventional TFBM

The principle of conventional TFBM is introduced in [3, 5] in detail. The main idea of TFBM is based on the equation (4) saying that one source is dominant in each TF slot. If we are able to estimate the dominant TF slots for each target signal $s_k(f,\tau)$ we can get a good estimation of the target signals $\hat{y}_k(f,\tau)$ by simply extracting the dominant TF slots from an arbitrary selected observation $x_P(f,\tau)$, $P \in \{1,...,M\}$.

TFBM estimates the dominant TF slots by normalization of the observation vector $\mathbf{x}(f,\tau)$ and clustering [5, 7]. Each cluster represents TF slots of an independent source. The estimated clusters are then used for building a TF binary mask $M_k(f,\tau)$. The extraction is then done by multiplying the TF binary mask $M_k(f,\tau)$ with observation $x_P(f,\tau)$. $M_k(f,\tau) = 1$ when a source k is estimated as dominant source in (f,τ), otherwise $M_k(f,\tau) = 0$. Due to the TF binary masking the TFBM outputs $\hat{y}_k(f,\tau)$ are distorted by musical noise.

Multi-channel TFBM

In our system, we use the TFBM as a pre-separation system and the final separation is done by multiple-beamformer. In order to design the multiple-beamformer the conventional TFBM must be extended:

1) Estimation of observation in the jammer only period $\hat{n}_1(f,\tau)$. This can be done by combining the pre-separated signals $\hat{y}_k(f,\tau)$ in TF domain. For example if N=3, M=3, P=1 and k=1 then the estimated target signal is $\hat{y}_1(f,\tau)$ and the estimated observation in jammer only period is $\hat{n}_1(f,\tau)=\hat{y}_2(f,\tau)+\hat{y}_3(f,\tau)$. These signals are extracted from the mixture $x_1(f,\tau)$.

2) Multi-channel extension. The extension can be achieved by multiplying the TF binary mask $M_k(f,\tau)$ with whole observation vector $\mathbf{x}(f,\tau)$. This results in vector of pre-separated target signal $\hat{\mathbf{y}}_k(f,\tau)=[\hat{y}_{k1}(f,\tau),\ldots,\hat{y}_{k\mathrm{M}}(f,\tau)]^{\mathrm{T}}$ and observation vector in jammer only period $\hat{\mathbf{n}}_k(f,\tau)=[\hat{n}_{k1}(f,\tau),\ldots,\hat{n}_{k\mathrm{M}}(f,\tau)]^{\mathrm{T}}$.

3) Estimation of mixing vector $\hat{\mathbf{h}}_k(f)$. We estimate $\hat{\mathbf{h}}_k(f)$ by back-normalization (inverse process to normalization of observation vector $\mathbf{x}(f,\tau)$) as described in [7].

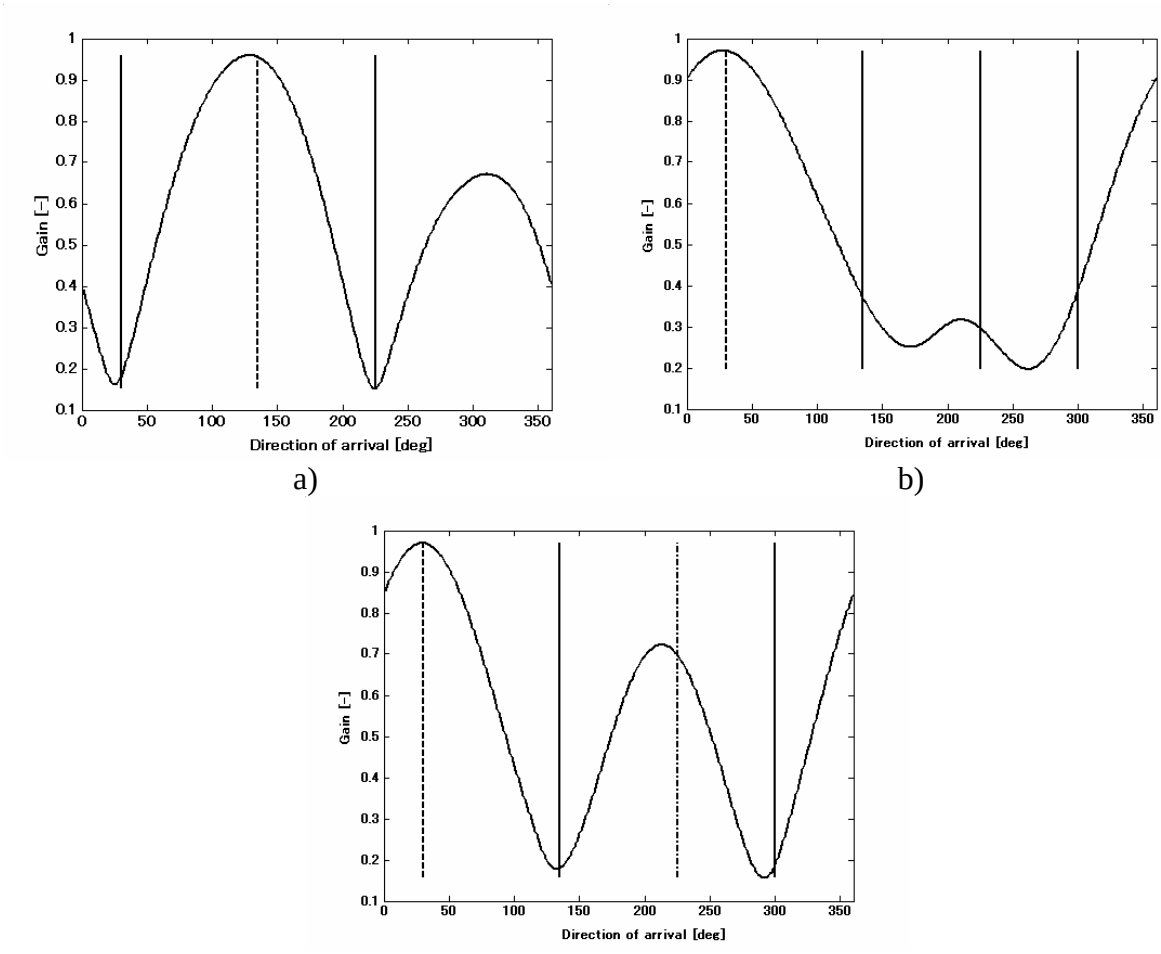

Figure 2: Spatial characteristic of ABF (vertical solid line is the direction of arrival (DOA) of jammer signals, dash line is the DOA of the target signal, dot-dashed line is the DOA of the excluded jammer) designed by our proposed system in a) DC, N=3, M=3, $\mathbf{R}_k(f)=\mathrm{E}[\hat{\mathbf{n}}_k(f,\tau)\hat{\mathbf{n}}_k(f,\tau)^{\mathrm{H}}]$; b) UDC, N=4, M=3, $\mathbf{R}_k(f)=\mathrm{E}[\hat{\mathbf{n}}_k(f,\tau)\hat{\mathbf{n}}_k(f,\tau)^{\mathrm{H}}]$ - without jammer reduction; c) UDC, N=4, M=3, $\mathbf{R}_k(f)=\mathrm{E}[\hat{\mathbf{n}}'_k(f,\tau)\hat{\mathbf{n}}'_k(f,\tau)^{\mathrm{H}}]$ - with jammer reduction.

3.3. Jammer reduction

Proposed approach shown in Fig. 1 can be used for both DC and UDC. The output of the TFBM and the Jammer reduction block form the input signal of ABF

$$\hat{\mathbf{x}}_k(f,\tau)=\hat{\mathbf{n}}'_k(f,\tau)+\hat{\mathbf{y}}_k(f,\tau). \quad (7)$$

In DC the block Jammer reduction does not affect the signal. Thus we can write $\hat{\mathbf{n}}_k(f,\tau)=\hat{\mathbf{n}}'_k(f,\tau)$ and $\hat{\mathbf{x}}_k(f,\tau)=\mathbf{x}(f,\tau)$. Now let us have a look at spatial characteristic of the ABF for DC in Fig. 2a. It can be seen that the jammers are suppressed well although the design is made blindly. However if we use the same approach in the UDC the jammers can not be suppressed efficiently as shown in Fig. 2b.

In order to fit the minima of spatial characteristics in the direction of arrival of the jammer signals we reduce the number of jammers in block Jammer reduction to M-1 jammers in UDC. For example N=4, M=3 and k=1 then $\hat{\mathbf{n}}_1(f,\tau)=\hat{\mathbf{y}}_2(f,\tau)+\hat{\mathbf{y}}_3(f,\tau)+\hat{\mathbf{y}}_4(f,\tau)$ and $\hat{\mathbf{n}}'_1(f,\tau)=\hat{\mathbf{y}}_3(f,\tau)+\hat{\mathbf{y}}_4(f,\tau)$. $\hat{\mathbf{y}}_2(f,\tau)$ was excluded from the vector $\hat{\mathbf{n}}'_1(f,\tau)$. The excluded signals are determined by the selection criterion. Different selection criteria can be used. In our approach we exclude the jammer signals with the fewest dominant TF slots in order to minimize musical noise. See Fig. 2c for spatial characteristic of ABF when jammer reduction is applied.

4. Experiments

We performed experiments for a DC (M=3, N=3) and an UDC (M=3, N=4) in a room with a reverberation time of 120 ms, see Fig. 3 for room setup. 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)

$$\mathrm{SIR}_k = 10\log_{10}\frac{\mathrm{E}\left[y_{kk}(t)^2\right]}{\mathrm{E}\left[\sum_{b=1,b\neq k}^{\mathrm{N}} y_{kb}(t)^2\right]}\,[\mathrm{dB}], \quad (8)$$

$$\mathrm{SDR}_k = 10\log_{10}\frac{\mathrm{E}\left[x_{\mathrm{P}k}(t)^2\right]}{\mathrm{E}\left[\left(x_{\mathrm{P}k}(t)-\gamma y_{kk}(t-\Delta)\right)^2\right]}\,[\mathrm{dB}], \quad (9)$$

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 $x_{k\mathrm{P}}(t)=\sum_l h_{\mathrm{P}k}(l)s_k(t-l)$. P is an arbitrarily selected sensor. Coefficients γ and Δ compensate the amplitude and delay, respectively.

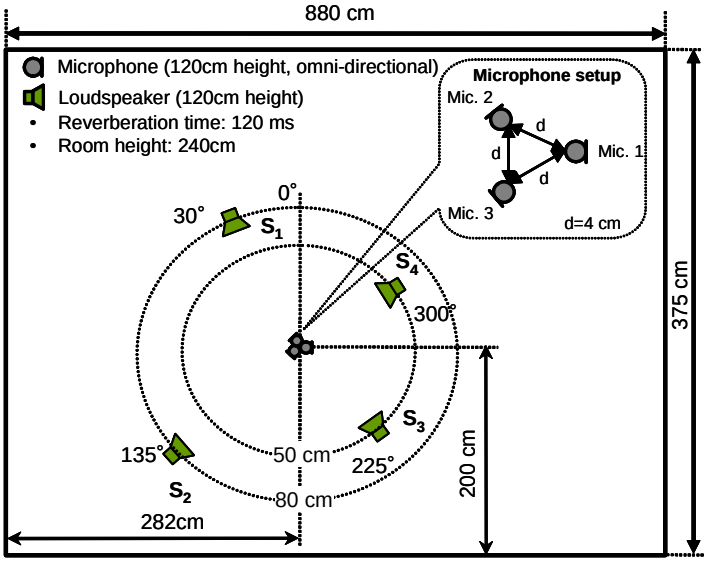

Figure 3: Room setup.

Table 2 shows the separation results for the DC, when sources S_1, S_2 and S_3 were used, and for the UDC. We used different BSS approaches in order to compare them. An conventional design ABF (CDABF), namely using steering vector with given target locations and observation correlation matrix $\mathbf{R}(f)$. Note that separation is *not* performed blindly. The conventional TFBM setup corresponds to the approach outlined in the beginning of section 3.2. Finally, we used our proposed method, which exploits $\hat{\mathbf{h}}(f)$, $\hat{\mathbf{R}}_k(f)$ for ABF design.
The CBABF 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 could not reflect the room reverberation or sensor misalignment. The TFBM achieved higher SIR and SDR values than the CBABF but the zero padding by the TF binary mask means that we hear large musical noise, especially in UDC. On the other hand, the proposed methods achieve higher separation performance with much less and without musical noise for UDC and DC, respectively.

Design Method	DC		UDC	
	SIR [dB]	SDR [dB]	SIR [dB]	SDR [dB]
CBABF	5.1	7.3	2.3	7.2
TFBM	10.9	10.6	9.6	9.0
Proposal	14.6	13.6	9.3	10.8

Table 2: Results for 3 mic. setting; input SIR=-3.1 dB in DC and input SIR=-4.8 dB in UDC.

5. Conclusion

In this paper, we have described a BSS approach applicable for both (over-)determined and underdetermined cases by assuming source sparseness. In (over-)determined case BSS is completely done by the multiple-beamformer, which does not cause musical noise. In underdetermined case the TFBM affect the multiple-beamformer input by reducing the number of jammers. Jammer reduction causes musical noise but in our approach just N-M jammers are reduced (zero padded) instead of N-1 jammers as in conventional TFBM approach and therefore the musical noise is significantly reduced. Furthermore our method provides better separation performance than conventional techniques.

Acknowledgement: The paper has been supported by the National Research Project "Information Society" 1ET301710509 and by the Ministry of Education, Youth and Sports, project number F2101/2006/61.

References

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

[2] S. Makino, H. Sawada, R. Mukai and S Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," IEICE Trans. Fundamentals, vol.E88-A, no.7, pp.1640-1655, July 2005.

[3] 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.

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

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

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

[7] 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.