GENERALIZED EIGENVECTOR BLIND SPEECH SEPARATION UNDER COHERENT NOISE IN A GSC CONFIGURATION

Dang Hai Tran Vu, Alexander Krueger, Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, Germany

{tran, krueger, haeb}@nt.uni-paderborn.de

ABSTRACT

This paper deals with a new technique for multi-channel separation of speech signals from convolutive mixtures under coherent noise. We demonstrate how the scaled transfer functions from the sources to the microphones can be estimated even in the presence of stationary coherent noise. The key to this are generalized eigenvalue decompositions of the power spectral density (PSD) matrices of the noisy speech and noise-only microphone signals with a controlled estimation of these matrices exploiting time-frequency sparseness of the speech sources. Separation is further improved by subsequent Gram-Schmidt orthogonalization which places spatial nulls at the interferers' directions, while noise reduction is improved by employing a novel blocking matrix and adaptive interference canceller in a Generalized Sidelobe Canceller (GSC)-like structure. We report promising experimental results for 2 speech sources with significant coherent noise in reverberant environments (RT60=0ms..500ms).

Index Terms— Noisy Source Separation, Sparse Signal Separation, Maximum SNR Beamformer, Speech Enhancement

1. INTRODUCTION

Blind source separation (BSS) is an approach for source signal reconstruction given an unknown mixture of source signals captured by the sensors. The BSS of speech signals has many applications including hands-free telecommunication or auditory scene analysis in a conference situation. In such applications one usually has to deal with sensor signals which are degraded by stationary noise of unknown spectral and spatial characteristics.

Methods for BSS have focused mainly on two different approaches: Independent Component Analysis (ICA) and Time-Frequency Masking (TFM). Techniques based on ICA usually involve higher order statistics and non-linear cost functions [1]. To deal with convolutive mixtures ICA can be employed bin-wise in frequency domain [2], however at the expense of permutation ambiguity. Time domain ICA approaches for convolutive mixtures have also been proposed [3] avoiding the permutation problem.

Another versatile approach is time-frequency masking (TFM) requiring approximately disjoint orthogonality of source signals [4]. This assumption holds for speech signals and reduces the BSS to a clustering problem of the observations [5]. TFM adds spectral subtraction principles to the source separation task but also inherits loss of speech quality and musical tones from these.

The focus of this paper is on multi-channel source separation in the presence of background noise, wherein noise-only periods are available and noise characteristics are stationary. While the consideration of noisy sensor signals is common in adaptive beamforming, it is often disregarded in the source separation literature. In [6] an algorithm for isotropic diffuse noise fields was presented. However, this method is based on the free field assumption and does not employ spatial filtering for source separation.

Recently we have proposed a blind beamforming technique based on the frequency-bin-wise generalized eigenvalue decomposition (GEVB) [8]. Using a single-channel postfilter it was able to approach the performance of the minimum variance distortionless response (MVDR) beamformer, however without requiring a priori information about the array geometry or the source-to-sensor transfer functions. In this paper we extend this concept in various ways to arrive at a BSS solution in the presence of additive noise. By exploiting time-frequency sparseness of speech we are able to estimate the transfer function ratios with a blind GEV beamforming approach even if multiple sources are active and even if noise is present at all times. To improve the suppression of the interfering source a Gram-Schmidt orthogonalization is applied to place spatial nulls at the interferer's direction. Permutation alignment is achieved by minimizing inter-frequency correlation of the output signals, similar to [11]. Finally, to improve the suppression of coherent noise, the transfer function ratio estimates are used to derive a blocking matrix (BM), which provides noise-only references for the adaptive interference canceller (AIC) in a Generalized Sidelobe Canceller (GSC) configuration. Figure 1 depicts the overall system architecture in case of 2 sources.

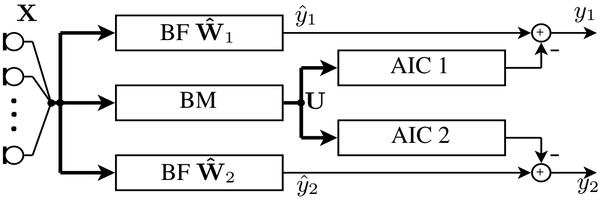

Fig. 1. *GSC-like system structure with beamformers (BFs), blocking matrix (BM) and adaptive interference cancellers (AICs)*

2. PROPOSED METHOD

We are given an array of M microphones and N Sources. With a K-point short-time Fourier Transform (STFT) a frequency representation of the signal at the j-th microphone is given by:

$$x_j(k,m) = \sum_{i=1}^{N} h_{ij}(k)s_i(k,m) + n_j(k,m) \quad j = 1..M \tag{1}$$

or in a more compact vector notation,

$$\mathbf{X}(k,m) = \sum_{i=1}^{N} \mathbf{H}_i(k)s_i(k,m) + \mathbf{N}(k,m) \tag{2}$$

where $k = 1, ..., K$ denotes the frequency bin and $m > 0$ is the time-frame index. $h_{ij}(k)$ are the transfer functions from the i-th source to the j-th microphone, $s_i(k,m)$ and $n_j(k,m)$ are STFTs of the source s_i and noise n_j respectively.

For sparse signals, such as speech, it holds that only a single source is present at any given time-frequency bin [4]. Then expression (2) can be approximated by

$$\mathbf{X}(k,m) \approx \mathbf{H}_i(k)s_i(k,m) + \mathbf{N}(k,m) =: \mathbf{X}_i(k,m) \qquad (3)$$

where, in a slight abuse of notation, the index i shall indicate the dominant source in time-frequency bin (k,m).

2.1. Adaptive Generalized Eigenvector Beamforming

For simplicity let us assume for a moment that only source s_i is active. The beamformer output for source s_i is given by

$$g_i(k,m) = \mathbf{F}_i^{\mathbf{H}}(k,m)\mathbf{X}(k,m) \qquad (4)$$

with the beamformer coefficient vector $\mathbf{F}_i(k,m)$. The design criterion for the Generalized Eigenvector Beamforming (GEVB) is to find beamformer coefficients $\mathbf{F}_i(k,m)$ which maximize the SNR in each frequency bin k:

$$\mathbf{F}_{i,\text{SNR}}(k) := \arg\max_{\mathbf{F}_i} \frac{\mathbf{F}_i^{\mathbf{H}}(k)\mathbf{\Phi}_{\mathbf{X}_i\mathbf{X}_i}(k)\mathbf{F}_i(k)}{\mathbf{F}_i^{\mathbf{H}}(k)\mathbf{\Phi}_{\mathbf{NN}}(k)\mathbf{F}_i(k)} - 1 \qquad (5)$$

where $\mathbf{\Phi}_{\mathbf{X}_i\mathbf{X}_i}(k) = \mathrm{E}[\mathbf{X}_i(k,m)\mathbf{X}_i^{\mathbf{H}}(k,m)]$ and $\mathbf{\Phi}_{\mathbf{NN}}(k) = \mathrm{E}[\mathbf{N}(k,m)\mathbf{N}^{\mathbf{H}}(k,m)]$ are short-time cross power spectral density matrices (PSDs) of noisy speech and noise respectively. $\mathbf{F}_{i,\text{SNR}}(k)$ have to be constrained to unit norm. Note that the PSD of the noisy speech is independent of the frame index m in our notation. This is not correct since the source signal is assumed to be nonstationary. Nevertheless it can be shown that the optimum solution is equal for all frame indices. Therefore we keep this simplified notation. It is shown in [8] that the optimum coefficient vector $\mathbf{F}_{i,\text{SNR}}(k)$ is the principal eigenvector of $\mathbf{\Phi}_{\mathbf{NN}}^{-1}(k)\mathbf{\Phi}_{\mathbf{X}_i\mathbf{X}_i}(k)$ and furthermore $\mathbf{F}_{i,\text{SNR}}(k)$ is related to the the transfer function vector $\mathbf{H}_i(k)$ by

$$\hat{\mathbf{H}}_i(k) := \mathbf{\Phi}_{\mathbf{NN}}(k)\mathbf{F}_{i,\text{SNR}}(k) = \zeta(k)\mathbf{H}_i(k) \qquad (6)$$

where $\zeta(k)$ is an arbitrary complex scalar.

Before solving the generalized eigenvalue problem the PSD matrices need to be determined. The estimation of $\mathbf{\Phi}_{\mathbf{NN}}$ can be easily be done in noise only periods, e.g. using an exponential time window

$$\begin{aligned}\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k,m) &= (1-\beta)\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k,m-1) \\ &+\beta(\mathbf{X}(k,m)\mathbf{X}^{\mathbf{H}}(k,m))|_{\mathbf{X}=\mathbf{N}} \qquad (7)\end{aligned}$$

with an appropriate initialization for $\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k,0)$ and $0 < \beta < 1$. To get an estimation for $\mathbf{\Phi}_{\mathbf{X}_i\mathbf{X}_i}(k)$ we can proceed in the same manner in speech plus noise periods:

$$\begin{aligned}\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m) &= (1-\alpha)\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m-1) \\ &+\alpha(\mathbf{X}(k,m)\mathbf{X}^{\mathbf{H}}(k,m))|_{\mathbf{X}=\mathbf{X}_i}. \qquad (8)\end{aligned}$$

where α is a time constant $0 < \alpha < 1$. Hence we need a voice activity detector (VAD) to discriminate between these two cases.

The estimation of the principal eigenvector can be carried out by using the power iteration method [9]:

$$\tilde{\mathbf{F}}_i(k,m) = \hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(k,m)\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m)\hat{\mathbf{F}}_i(k,m-1) \qquad (9)$$

$$\hat{\mathbf{F}}_i(k,m) = \frac{\tilde{\mathbf{F}}_i(k,m)}{\left\|\tilde{\mathbf{F}}_i(k,m)\right\|} \qquad (10)$$

This simple algorithm showed excellent convergence behavior and good estimates for $\mathbf{F}_{i,\text{SNR}}(k)$ and thus for the scaled transfer function $\hat{\mathbf{H}}_i(k)$ in practical tests.

2.2. Separation procedure

Now we turn back to the multi-speaker scenario with N simultaneously active sources. Based on the sparse source assumption (3) it is obviously possible to estimate all transfer functions $\hat{\mathbf{H}}_i(k)$ if we update the PSD matrices $\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m)$ for every source only in time-frequency bins where the source s_i is dominant. A simple modification of equation (8) accounts for this consideration:

$$\begin{aligned}\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m) &= (1-\alpha b_i(k,m))\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m-1) \\ &+\alpha b_i(k,m)(\mathbf{X}(k,m)\mathbf{X}^{\mathbf{H}}(k,m)) \qquad (11)\end{aligned}$$

where $b_i(k,m)$ is a binary mask typically defined in sparse source BSS approaches [4]:

$$b_i(k,m) = \begin{cases} 1, & \text{if source } s_i \text{ is dominant} \\ 0, & \text{else.} \end{cases} \qquad (12)$$

Unfortunately, reliable estimates of $b_i(k,m)$ are hard to obtain in a reverberant environment. Consequently we fall back to a soft decision:

$$b_i(k,m) = \gamma(l_i(k,m)) \qquad (13)$$

where $l_i(k,m)$ is the source activity likelihood and $\gamma(\cdot)$ is a nonlinear decision function. We employ the following soft decision function:

$$\gamma(u) = \tanh\left((u + c_1)^{c_2}\right) \qquad (14)$$

In figure 2 an example of the decision characteristic of equation (14) is given for some appropriate parameters c_1 and c_2.

To obtain an estimate of $l_i(k,m)$ we propose a feedback loop by using the power ratio at the beamformer outputs:

$$l_i(k,m) = \frac{|g_i(k,m)|^2}{\sum_{n=1}^{N}|g_n(k,m)|^2}. \qquad (15)$$

Since $\mathbf{F}_i(k,m)$ is normalized, see (10), equation (15) can be seen as an input vector matching score. Thus this feedback in combination with the adaptive GEVB technique results in a straightforward observation vector clustering algorithm. This is similar to [5] however with special consideration for spatially correlated noise.

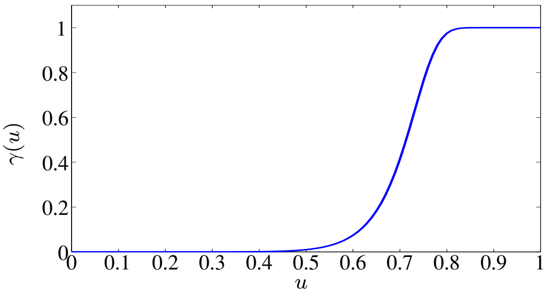

Fig. 2. *Soft decision function* ($c_1 = 0.25, c_2 = 16$).

2.3. Permutation alignment

Since the separation is carried out in each frequency bin separately we are suffering from arbitrary permutation of the sources in each frequency bin. In order to reconstruct properly separated speech signals in time-domain, frequency-domain separated signals originating from the same source should be aligned together. Since this well known problem in frequency domain blind source separation is not the focus of this paper we refer to a recently proposed method [11].

2.4. Gram-Schmidt Orthogonalization and Blocking Matrix

Computing the system output as $y_{i,\mathrm{MF}}(k,m) = \hat{\mathbf{H}}_i^{\mathbf{H}}(k)\mathbf{X}(k,m)$ with $\hat{\mathbf{H}}_i(k) = \mathbf{\Phi}_{\mathbf{NN}}(k)\hat{\mathbf{F}}_i(k)$ corresponds to spatial matched filtering. While this will form a spatial pattern with main lobe in the direction of the i-th source, no special spatial suppression is achieved regarding the interferers. Thus the output has a good speech quality but contains also strong interfering signals. To gain in signal-to-interference (SIR) we apply a Gram-Schmidt orthogonalization to force mutual orthogonality of the filter coefficients. In the case of 2 sources this results in:

$$\mathbf{W}_{1/2}(k) := \left(\mathbf{I} - \frac{\hat{\mathbf{H}}_{2/1}(k)\hat{\mathbf{H}}_{2/1}^{\mathbf{H}}(k)}{\hat{\mathbf{H}}_{2/1}^{\mathbf{H}}(k)\hat{\mathbf{H}}_{2/1}(k)}\right)\hat{\mathbf{H}}_{1/2}(k). \qquad (16)$$

Likewise the BM for the GSC structure can easily obtained by

$$\mathbf{B}^{\mathbf{H}}(k) := \mathbf{I} - \frac{\hat{\mathbf{H}}_1(k)\hat{\mathbf{H}}_1^{\mathbf{H}}(k)}{\hat{\mathbf{H}}_1^{\mathbf{H}}(k)\hat{\mathbf{H}}_1(k)} - \frac{\mathbf{W}_2(k)\mathbf{W}_2^{\mathbf{H}}(k)}{\mathbf{W}_2^{\mathbf{H}}(k)\mathbf{W}_2(k)}. \qquad (17)$$

It can easily be verified that the noise reference signals $\mathbf{U}(k,m) = \mathbf{B}^{\mathbf{H}}(k)\mathbf{X}(k,m)$ do not contain any source signal components, if $\hat{\mathbf{H}}_{2/1}(k)$ are perfect estimates. A similar BM for a GSC-like structure in case of one source was also proposed in [10].

2.5. TDOA Estimation

While equation (16) place spatial nulls at the interferer direction, the filter coefficients have no constraint for target direction gain. If the room transfer function is given, a solution is achieved by postulating a distortionless response in target signal direction:

$$\hat{\mathbf{W}}_i(k) = \frac{\mathbf{W}_i(k)}{\mathbf{H}_i^{\mathbf{H}}(k)\mathbf{W}_i(k)} \qquad (18)$$

Since $\mathbf{H}_i(k)$ is not available, we approximate this unknown transfer function by

$$\mathbf{H}_i(k) \approx [1, e^{-j\omega_k\tau_{i,2}}, ..., e^{-j\omega_k\tau_{i,M}}]^{\mathbf{T}} \qquad (19)$$

where $\tau_{i,j}$ is the time difference of arrival (TDOA) between the the first and j-th sensor for the i-th source. Note that in absence of reverberation (19) becomes an equation.
We can obtain estimates for τ_i by searching for the maximum of the cross correlation of the impulse responses of the first and the j-th estimated room transfer function, where the correlation is typically carried out in the frequency domain:

$$\hat{\tau}_{i,j} = \arg\max_{\tau} \mathrm{IFFT}\left\{\left[\hat{H}_{i,1}(0)\hat{H}_{i,j}^*(0), ..., H_{i,1}(K)\hat{H}_{i,j}^*(K)\right]\right\}. \qquad (20)$$

With this normalization the output of the beamformer becomes finally $\hat{y}_i(k,m) = \hat{\mathbf{W}}_i^{\mathbf{H}}(k)\mathbf{X}(k,m)$. In [8] further variation possibilities for this normalization were proposed.

2.6. Method summary

Summing these considerations the algorithm becomes:

1. Use VAD to discriminate between noise-only and speech-presence periods.
2. Estimate $\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k)$ with equation (7) in noise-only periods.
3. Set $\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,0)$ with appropriate $\hat{\mathbf{F}}_i(k,0)$ to random values.
4. In every frame of speech-presence periods:
   (a) Compute intermediate output (4).
   (b) Calculate soft masking with (13) and (15).
   (c) Update $\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m)$ by applying (11).
   (d) Carry out one step of power iteration (9) and (10).
5. Use (6) to calculate $\hat{\mathbf{H}}_i(k)$.
6. Solve permutation alignment with [11].
7. Use Gram-Schmidt process to place spatial zeros (16) and build BM with (17).
8. Use TDOA estimates for proper gain factor.
9. Calculate Beamformer output $\hat{y}_i$ and noise references $\mathbf{U}$.
10. Conduct AIC adaptation in noise-only periods.
11. Calculate system output y_i.

3. EXPERIMENTS

In this section we experimentally evaluate the proposed blind source separation method for the case of two simultaneously active sources with a correlated noise source in a reverberant enclosure of size (6m) x (4m) x (2m). A uniform circular array (0.1m radius) with 8 microphones was used. The sources were positioned around the microphone array in 5 different locations. 10 utterances from different speakers (5 male and 5 female), sampled at 16 kHz, were used as source speech signals. Source signal durations are about 5 s. Taking 2 out of 10 utterances at 5 possible positions and 8 analyzed reverberation times between 0 ms and 500 ms results in 3600 audio files. Recordings of the fan noise of a video projector were used as coherent noise. The input power ratio of the two sources and the coherent noise was about 0 dB. To every microphone white noise with SNR of about 25 dB was added. The STFT frame size was 512 samples with an 1/4 shift. The AIC filter length was 1024 samples.
The system performance was evaluated in terms of signal-to-interference-ratio (SIR), signal-to-noise-ratio (SNR) and signal-to-distortion-ratio (SDR)

$$\mathrm{SIR} := 10\log_{10}\left(\frac{\mathrm{E}[\hat{s}^2(t)]}{\mathrm{E}[\bar{s}^2(t)]}\right)\,[\mathrm{dB}] \qquad (21)$$

$$\mathrm{SNR} := 10\log_{10}\left(\frac{\mathrm{E}[\hat{s}^2(t)]}{\mathrm{E}[\bar{n}^2(t)]}\right)\,[\mathrm{dB}] \qquad (22)$$

$$\mathrm{SDR} := 10\log_{10}\left(\frac{\mathrm{E}[\hat{s}^2(t)]}{\mathrm{E}[(\hat{s}(t) - a\hat{s}_{\mathrm{DSB}}(t-\delta))^2]}\right)\,[\mathrm{dB}] \qquad (23)$$

where $\hat{s}(t)$ is the time domain target signal component, $\bar{s}(t)$ is the interferer's component, $\bar{n}(t)$ is the noise component at the system output. The reference $\hat{s}_{\mathrm{DSB}}(t)$ for the speech distortion measurement was the output of a delay-and-sum Beamformer (DSB), whose optimal delay τ_j was assumed to be perfectly known, and where the parameters a and δ were chosen to maximize SDR. Thus the coefficients a and δ compensate the amplitude and delay.

Figure 3 shows the simulation results. For comparison we also apply our method if all PSD matrices $\mathbf{\Phi}_{\mathbf{X}_i\mathbf{X}_i}(k)$ and $\mathbf{\Phi}_{\mathbf{NN}}(k)$ are perfectly known a priori. Furthermore the performance is given with and without AIC path to gain insight into the noise suppression capabilities of the GSC-like structure.

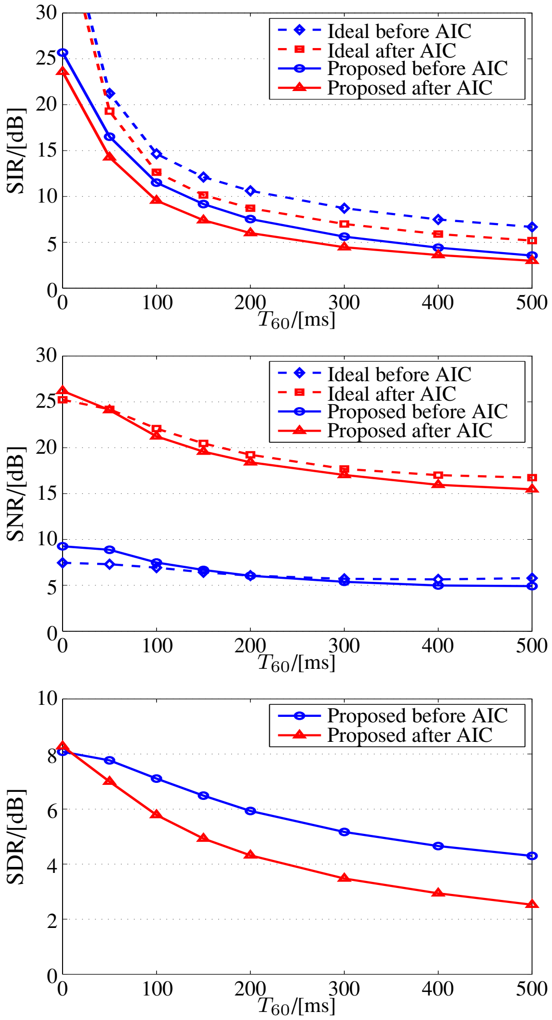

Fig. 3. *SIR, SNR and SDR of proposed method compared to ideal case of perfectly known PSD matrices*

We achieved good separation results in low reverberation conditions. As expected separation performance of the proposed method decreases for higher reverberation times. The offset to the ideal is caused by a model mismatch since equation (3) is an approximation and does not reflect the mixing situation correctly.

Noise suppression is very large at low reverberation times and surprisingly good even at high reverberation times. Due to mismatch of transfer function estimates and limited filter lengths we are suffering from leakage of source signals into the AIC path, which results in a somewhat lower SIR after noise cancellation. Hence we have to sacrifice a little bit separation performance for a significant boost in noise suppression.

Speech quality evaluation gives satisfying results in low reverberation conditions. At high reverberation time the SDR measurement has to be viewed with caution since a fair quantitative comparison especially in reverberant environment is difficult. In hearing tests speech quality with and without AIC is hardly distinguishable from each other.

4. CONCLUSIONS

A blind speech separation method with special account for spatially correlated noise under sparse source assumption has been presented. Our method corresponds to a blind system identification and filter synthesizing approach. We also confirmed that the proposed algorithm works well in low to medium reverberation environment.

5. ACKNOWLEDGMENTS

This work was in part supported by Deutsche Forschungsgemeinschaft under contract number HA 3455/4-1.

6. REFERENCES

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

[2] H. Sawada, R. Mukai, S. Araki, S. Makino, “Frequency domain blind source separation”, in *Speech Enhancement*, Springer, 2005.

[3] H. Buchner, R. Aichner, W. Kellermann, “A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics”, in *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, pp. 120 - 134, 2005.

[4] O. Yilmaz, S. Richard, “Blind separation of speech mixtures via time-frequency masking”, in *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830-1847, 2004.

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

[6] R. Balan, J. Rosca, S. Rickard, “Non-Square blind source separation under coherent noise by beamforming and time-frequency masking”, in *Proc. ICA2005*, 2003.

[7] S. Araki, H. Sawada, S. Makino, ”Blind speech separation in a meeting situation with maximum SNR beamformers”, in *Proc. ICASSP2007*, 2007.

[8] E. Warsitz, R. Haeb-Umbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition”, in *IEEE Trans. Audio Speech Language Processing*, vol. 15, no. 5, pp. 1529-1539, 2007.

[9] J. Karhunen, “Adaptive algorithms for estimating eigenvectors of correlation type matrices”, in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 9, no. 9, pp. 592-595, 1984.

[10] E. Warsitz, A. Krueger, R. Haeb-Umbach, “Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller”, in *Proc. ICASSP2008*, 2008.

[11] H. Sawada, S. Araki and S. Makino, “Measuring dependence of bin-wise separated signals for permutation alignment in Frequency-domain BSS”, in *Proc. ISCAS2007*, 2007.