Blind Speech Separation in Presence of Correlated Noise with Generalized Eigenvector Beamforming

Dang Hai Tran Vu, Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn,
E-Mail: {tran,haeb}@nt.uni-paderborn.de

Abstract

This paper considers the convolutive blind source separation of speech sources in the presence of spatially correlated noise. We introduce a method for estimating the scaled mixing matrix from the sources to the microphones even if coherent noise is present. This is achieved by combining time-frequency sparseness with the generalized eigenvalue decomposition of the power spectral density matrix (PSD) of the noisy speech and noise-only microphone signals. Separation is performed by spatial filtering with coefficients constructed by Gram-Schmidt orthogonalization which places spatial nulls at the interferer's direction. Experimental results show that our approach is capable of separating 2 sources in a reverberant environment (RT60=0ms..500ms) degraded by significant directional noise.

1 Introduction

The objective of blind source separation (BSS) is to extract source signals from mixed signal information observed at each sensor and, possibly, to estimate the unknown mixing channel. The BSS technique for speech dealt with in this paper can be used in many applications of speech enhancement including hands-free telecommunication and automatic meeting note taking.

In recent source separation literature two main approaches have emerged. One is based on independent component analysis (ICA) and the other relies on the sparseness of source signals.

ICA techniques can be applied when the number of sources N is equal or less than the number of sensors M ($N \leq M$). These methods usually involve higher order statistics and non-linear cost functions. They can be carried out in time domain or in frequency domain. In time-domain ICA separation is conducted directly on the convolutive mixtures and unmixing FIR filters are estimated directly [1]. However the convergence of most time-domain ICA algorithms is slow since the adaptation of long filters in time-domain is very complex and computationally demanding. By contrast, in frequency-domain ICA mixtures are transformed to frequency domain and instantaneous ICA is applied in each frequency bin separately [2]. While computational complexity is greatly reduced frequency-domain algorithms suffer from arbitrary permutation and scaling in each bin, which leads to large signal distortions. Since these problems are inherent to frequency-domain BSS approaches, various solutions have been proposed [3] [4] where the computational advantages of frequency-domain approaches are diminished by the effort required for solving the permutation and scaling problem.

By contrast sparseness based methods can separate all sources even when $N > M$. These methods can be divided into two main categories. One extracts the source signals with binary time-frequency masks [5]. This has the advantage of having low computational requirements at the expense of signal quality impairments and musical tones. The other category of methods recover the sources after mixing matrix estimation, which avoid these issues, however in underdetermined BSS the mixing matrix cannot simply be inverted [6]. The method discussed in this paper belongs the latter category of sparseness based BSS approaches.

In realistic scenarios the microphone signals are degraded by stationary noise of unknown spectral and spatial characteristics. At high signal-to-noise ratios (SNR) one can neglect this problem and rely on the robustness of the employed BSS technique. If the noise level is significant BSS approaches derived with particular consideration of noise have to be applied. Unfortunately consideration of coherent noise is usually disregarded in the source separation literature. Consequently techniques for noisy BSS are rarely developed and are very scarcely studied in practical scenarios. It can be shown that additive noise introduces a bias in the unmixing filter coefficients. In [7] bias removal for adaptive ICA Algorithms are discussed but these technique have poor convergence behavior and so far have only been applied to instantaneous ICA tasks with artificial mixed sources. In [8] an algorithm for isotropic diffuse noise fields was presented. However this method is based on the free field assumption and employs source signals reconstruction with binary time-frequency masks.

In contrast to BSS literature, consideration of noisy sensors signals is common in adaptive beamforming. Recently a blind beamforming technique based on the frequency-binwise generalized eigenvalue decomposition (GEVB) have been proposed [9]. 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 [3].

2 Proposed Method

Let $s_1(t),...,s_N(t)$ be the desired speech sources to be recovered from the convolutive mixtures $x_1(t),...,x_M(t)$ given by

$$x_j(t) = \sum_{i=1}^{N} \sum_{l} h_{ij}(l) s_i(t-l) + n_j(t) \quad j = 1,...,M \qquad (1)$$

where $h_{ij}(l)$ is the impulse response from source i to microphone j, and $n_j(t)$ is the observed stationary noise at sensor j. With the K-point short-time Fourier Transform (STFT) equation (1) can be converted into

$$x_j(k,m) = \sum_{i=1}^{N} H_{ij}(k)S_i(k,m) + N_j(k,m) \quad j = 1,...,M \quad (2)$$

or in a more compact vector notation,

$$\mathbf{X}(k,m) = \sum_{i=1}^{N} \mathbf{H}_i(k)S_i(k,m) + \mathbf{N}(k,m) \quad (3)$$

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 $s_i(t)$ and noise $n_j(t)$ respectively.
For sparse signals, such as speech, it holds that only a single source is present at any given time-frequency bin [5]. Then expression (3) can be approximated by

$$\mathbf{X}(k,m) \approx \mathbf{H}_i(k)S_i(k,m) + \mathbf{N}(k,m) =: \mathbf{X}_i(k,m) \quad (4)$$

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

$$g_i(k,m) = \mathbf{F}_i^{\mathbf{H}}(k,m)\mathbf{X}(k,m) \quad (5)$$

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:

$$\mathbf{F}_{i,\mathrm{SNR}}(k) := \arg\max_{\mathbf{F}_i} \frac{\mathbf{F}_i^{\mathbf{H}}(k)\mathbf{\Phi}_{\mathbf{X}_i\mathbf{X}_i}(k)\mathbf{F}_i(k)}{\mathbf{F}_i^{\mathbf{H}}(k)\mathbf{\Phi}_{\mathbf{NN}}(k)\mathbf{F}_i(k)} - 1 \quad (6)$$

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. $\mathbf{F}_{i,\mathrm{SNR}}(k)$ 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 [9] that the optimum coefficient vector $\mathbf{F}_{i,\mathrm{SNR}}(k)$ is the principal eigenvector of $\mathbf{\Phi}_{\mathbf{NN}}^{-1}(k)\mathbf{\Phi}_{\mathbf{X}_i\mathbf{X}_i}(k)$ and furthermore $\mathbf{F}_{i,\mathrm{SNR}}(k)$ is related to the the transfer function vector $\mathbf{H}_i(k)$ by

$$\hat{\mathbf{H}}_i(k) := \mathbf{\Phi}_{\mathbf{NN}}(k)\mathbf{F}_{i,\mathrm{SNR}}(k) = \zeta(k)\mathbf{H}_i(k) \quad (7)$$

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

$$\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k,m) = (1-\beta)\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k,m-1) + \beta(\mathbf{X}(k,m)\mathbf{X}^{\mathbf{H}}(k,m))|_{\mathbf{X}=\mathbf{N}} \quad (8)$$

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:

$$\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m) = (1-\alpha)\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m-1) + \alpha(\mathbf{X}(k,m)\mathbf{X}^{\mathbf{H}}(k,m))|_{\mathbf{X}=\mathbf{X}_i}. \quad (9)$$

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 [10]:

$$\tilde{\mathbf{F}}_i(k,m) = \hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(k,m)\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m)\hat{\mathbf{F}}_i(k,m-1) \quad (10)$$

$$\hat{\mathbf{F}}_i(k,m) = \frac{\tilde{\mathbf{F}}_i(k,m)}{\left\|\tilde{\mathbf{F}}_i(k,m)\right\|} \quad (11)$$

This simple algorithm showed excellent convergence behavior and good estimates for $\mathbf{F}_{i,\mathrm{SNR}}(k)$ 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 (4) it is obviously possible to estimate the scaled mixing matrix consisting of 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 (9) accounts for this consideration:

$$\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m) = (1-\alpha b_i(k,m))\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m-1) + \alpha b_i(k,m)(\mathbf{X}(k,m)\mathbf{X}^{\mathbf{H}}(k,m)) \quad (12)$$

where $b_i(k,m)$ is a binary mask typically defined in sparse source BSS approaches [5]:

$$b_i(k,m) = \begin{cases} 1, & \text{if source } s_i \text{ is dominant} \\ 0, & \text{else.} \end{cases} \quad (13)$$

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

$$b_i(k,m) = \gamma(l_i(k,m)) \quad (14)$$

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

$$\gamma(u) = \tanh\left((u+c_1)^{c_2}\right) \quad (15)$$

In figure 1 an example of the decision characteristic of equation (15) 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:

$$l_i(k,m) = \frac{|g_i(k,m)|^2}{\sum_{n=1}^{N}|g_n(k,m)|^2}. \quad (16)$$

Since $\mathbf{F}_i(k,m)$ is normalized, see (11), equation (16) 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 concept is well known in literature however with special consideration for spatially correlated noise.

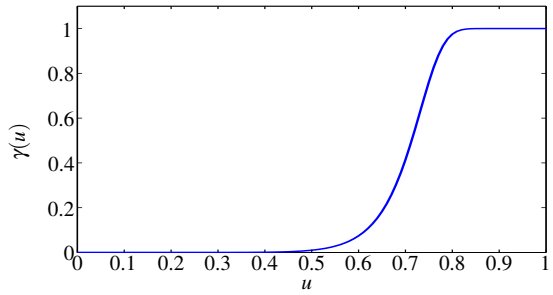

Figure 1: *Soft decision function* ($c_1 = 0.25, c_2 = 16$).

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 [3].

2.4 Gram-Schmidt Orthogonalization

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:

$$\mathbf{W}_{1/2}(k) := \left(\mathbf{I} - \frac{\hat{\mathbf{H}}_{2/1}(k)\hat{\mathbf{H}}_{2/1}^{\mathbf{H}}(k)}{\hat{\mathbf{H}}_{2/1}^{\mathbf{H}}(k)\hat{\mathbf{H}}_{2/1}(k)}\right)\hat{\mathbf{H}}_{1/2}(k). \quad (17)$$

Note that this approach is not applicable in underdetermined separation tasks ($N > M$) since the dimensionality of the orthogonal subspace is limited by $M-1$. In that case demixing filters can be obtained by maximum a posteriori estimation proposed in [6] or we have to fall back to binary masking for source reconstruction [5].

2.5 TDOA Estimation

While equation (17) 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:

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

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

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

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:

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

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 [9] alternatives 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 (8) 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 (5).
   (b) Calculate soft masking with (14) and (16).
   (c) Update $\hat{\mathbf{\Phi}}_{\mathbf{X}_i\mathbf{X}_i}(k,m)$ by applying (12).
   (d) Carry out one step of power iteration (10) and (11).
5. Solve permutation alignment with [3].
6. Use Gram-Schmidt process to place spatial zeros (17).
7. Use TDOA estimates for proper gain factor.
8. Calculate system output $\hat{y}_i$.

3 Simulation Results

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 at the level of -25dB below signal power 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)

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

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

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

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 where chosen to maximize SDR. Thus the coefficients a and δ compensate the amplitude and delay.
Figure 2 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.

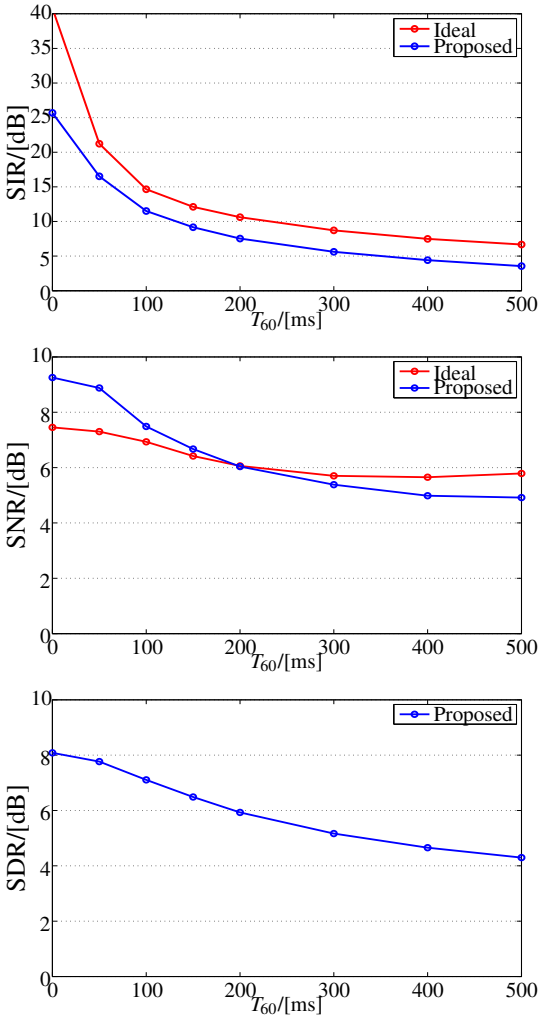

Figure 2: *SIR, SNR and SDR of proposed method compared to ideal case of perfectly known PSD matrices*

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 (4) is an approximation and does not reflect the mixing situation correctly.
As the focus of this paper is to demonstrate that source separation is possible even if significant coherent noise is present no special precautions regarding noise suppression are taken. Hence noise suppression results are low and is mostly achieved by inherent directivity properties of spatial filtering.
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 subjective hearing tests speech quality is quite satisfactorily.

4 Conclusions

A blind speech separation method with special account for correlated noise under the assumption of sparse sources has been presented. We confirmed that the proposed algorithm works well in low to medium reverberation environment. Future work will address further enhancement by supressing the correlated noise in a general sidelobe canceller (GSC) configuration.

5 Acknowledgements

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

References

[1] 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, 2005.

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

[3] 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.

[4] H. Sawada, R. Mukai, S. Araki and S. Makino, "A robust approach to the permutation problem of frequency-domain blind source separation", in *Proc. ICASSP2003*, 2003.

[5] 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.

[6] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and L1-norm minimization", in *EURASIP Journal on Advances in Signal Processing*, 2007.

[7] A. Cichocki, Shun-ichi Amari *Adaptive Blind Signal and Image Processing*, John Wiley and Sons, 2003

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

[9] 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.

[10] 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.