SPEECH ENHANCEMENT WITH A NEW GENERALIZED EIGENVECTOR BLOCKING MATRIX FOR APPLICATION IN A GENERALIZED SIDELOBE CANCELLER

Ernst Warsitz, Alexander Krueger and Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn,
33098 Paderborn, Germany
{warsitz, krueger, haeb}@nt.uni-paderborn.de

ABSTRACT

The generalized sidelobe canceller by Griffith and Jim is a robust beamforming method to enhance a desired (speech) signal in the presence of stationary noise. Its performance depends to a high degree on the construction of the blocking matrix which produces noise reference signals for the subsequent adaptive interference canceller. Especially in reverberated environments the beamformer may suffer from signal leakage and reduced noise suppression. In this paper a new blocking matrix is proposed. It is based on a generalized eigenvalue problem whose solution provides an indirect estimation of the transfer functions from the source to the sensors. The quality of the new generalized eigenvector blocking matrix is studied in simulated rooms with different reverberation times and is compared to alternatives proposed in the literature.

Index Terms— Speech enhancement, array signal processing

1. INTRODUCTION

The enhancement of a desired speech signal in the presence of stationary noise using an array of microphones has been studied for several years. A very famous and robust beamforming method is the generalized sidelobe canceller (GSC) [1] which consists of three signal processing units, see Fig. 1. While the fixed beamformer (FB) is designed to deliver a first estimate of the desired speech signal, the blocking matrix (BM) is supposed to block the speech signal parts in the microphone signals. The noise references at its output drive a multichannel adaptive interference canceller (AIC) whose coefficients are adapted to suppress the remaining noise in the FB output.

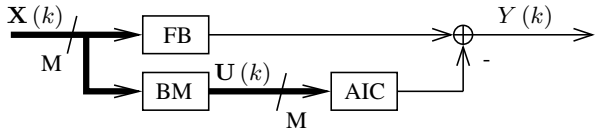

Fig. 1. *GSC structure*

In [1] the sound propagation from source to sensors is characterized by pure delays. Hence, noise reference signals can be obtained by pairwise subtraction of time-aligned microphone signals. The alignment requires knowledge of the direction of arrival (DoA) of the speech signal and of the array geometry. However, in real acoustic environments multipath propagation from the source to the sensors causes leakage of speech signal components into the noise references, resulting in a reduced noise suppression and distortion of the desired signal.

This work is partially supported by the DFG RTG GK-693 of the Paderborn Institute for Scientific Computation (PaSCo) and project Ha 3455/4-1.

Hoshuyama et al. proposed an adaptive blocking matrix (ABM), where the optimization criterion for adaptation is to achieve noise reference signals orthogonal to the FB output signal. They furthermore introduced constraints on the ABM filter coefficients to improve the robustness against estimation errors of the DoA [2]. Herbordt and Kellerman developed an efficient frequency domain realization for the GSC with ABM [3]. They showed that for optimal noise reference signals the adaptation has to be carried out only in periods of absence of noise [4]. In situations where this condition does not hold, only suboptimal solutions can be achieved.

In order to cope with the problem of stationary noise Gannot et al. [5] introduced the transfer function ratios blocking matrix (TFRBM). The ratios of the transfer functions (TF) from the speaker to the mirophones are estimated with the least squares method in periods when the speech signal is present.

Here, we propose a new blocking matrix which is based on the idea of generating a speech reference signal similar to [2]. But opposite to [2] we first use statistically optimal beamformer filter coefficients resulting from maximizing the output signal-to-noise ratio. Secondly the orthogonal projection for constructing the blocking matrix is done directly without LMS adaptation. The optimal filter coefficients are computed iteratively by solving a generalized eigenvalue problem (GEVP). With the proposed method a reduced signal distortion and a higher noise reduction compared to the TFRBM can be achieved.

2. GENERALIZED EIGENVECTOR BEAMFORMING

Consider an array of M microphones which is located in a reverberant enclosure. Each time-discrete microphone signal $x_i(l), (i = 1, ..., M)$, where l denotes the discrete-time index, is assumed to consist of two components: a signal component $s_i(l)$ which results from the convolution of the desired (speech) source signal $s_0(l)$ with the room impulse response $h_i(l)$ from the source position to the i-th microphone, and a stationary noise term $n_i(l)$:

$$x_i(l) = s_i(l) + n_i(l) = s_0(l) * h_i(l) + n_i(l). \quad (1)$$

The discrete Fourier transforms (DFTs) of the time signals are denoted by corresponding capital letters such as $X_i(k)$, where k means the frequency bin. For a block by block processing the microphone signals are windowed by a discrete window function $g(l)$ of a finite length L:

$$x_i(l, m) := x_i(l)g(l - (m - 1)B), \quad (2)$$

where B denotes the advance between successive blocks and m is the block index. The short time discrete Fourier transforms (STDFTs) of $x_i(l)$, denoted by $X_i(k, m)$, are arranged in a vector:

$$\mathbf{X}(k, m) := (X_1(k, m), ..., X_M(k, m))^T, \quad (3)$$

where $(\cdot)^T$ denotes transposition.

2.1. Maximum-SNR-criterion

For the blocking matrix design we use information about the transfer functions from the instationary signal source to the microphones. This information can be obtained from the optimal transfer functions $\mathbf{F}_{\text{SNR}}(k)$ of a generalized eigenvector beamformer [6] which results from the maximum-SNR-criterion

$$\mathbf{F}_{\text{SNR}}(k,m) := \underset{\mathbf{F}(k,m)}{\operatorname{argmax}} \operatorname{SNR}(k,m), \tag{4}$$

with the frequency dependent SNR for the m-th block

$$\operatorname{SNR}(k,m) := \frac{\mathbf{F}^{\mathbf{H}}(k,m)\,\mathbf{\Phi}_{\mathbf{SS}}(k,m)\,\mathbf{F}(k,m)}{\mathbf{F}^{\mathbf{H}}(k,m)\,\mathbf{\Phi}_{\mathbf{NN}}(k,m)\,\mathbf{F}(k,m)}. \tag{5}$$

The short time cross power spectral density matrices (PSDs) of speech and noise are given by

$$\mathbf{\Phi}_{\mathbf{SS}}(k,m) := E\left[\mathbf{S}(k,m)\,\mathbf{S}^{\mathbf{H}}(k,m)\right] \tag{6}$$

$$\mathbf{\Phi}_{\mathbf{NN}}(k,m) := E\left[\mathbf{N}(k,m)\,\mathbf{N}^{\mathbf{H}}(k,m)\right] = \mathbf{\Phi}_{\mathbf{NN}}(k), \tag{7}$$

where $(\cdot)^H$ denotes the conjugated transposition. The expectation is conducted over all realizations of the signals in the m-th block. The dependence of $\mathbf{\Phi}_{\mathbf{SS}}(k,m)$ on the block index m is obvious as speech signals are instationary. On the contrary the noise is here assumed to be stationary so that the block index for the corresponding PSD can be dropped. We further assume that the speech and noise are uncorrelated and that each of the signals has zero mean. This allows to split the PSD of the microphone signals into two parts

$$\mathbf{\Phi}_{\mathbf{XX}}(k,m) = \mathbf{\Phi}_{\mathbf{SS}}(k,m) + \mathbf{\Phi}_{\mathbf{NN}}(k). \tag{8}$$

In that case the solution of (4) is equivalent to the eigenvector belonging to the largest eigenvalue of the GEVP [6]

$$\mathbf{\Phi}_{\mathbf{XX}}(k,m)\,\mathbf{F}(k,m) = \lambda(k,m)\,\mathbf{\Phi}_{\mathbf{NN}}(k)\,\mathbf{F}(k,m). \tag{9}$$

With the assumption that $\mathbf{\Phi}_{\mathbf{NN}}(k)$ is not singular the GEVP can be transformed to the special eigenvalue problem (SEVP)

$$\mathbf{\Phi}_{\mathbf{NN}}^{-\mathbf{1}}(k)\,\mathbf{\Phi}_{\mathbf{XX}}(k,m)\,\mathbf{F}(k,m) = \lambda(k,m)\,\mathbf{F}(k,m). \tag{10}$$

The TFs from the desired source to the sensors are assumed to change slowly in time. Then for large window length L, the approximation

$$\mathbf{X}(k,m) \approx S_0(k,m)\,\mathbf{H}(k) + \mathbf{N}(k,m) \tag{11}$$

results from (1). It follows from above that the PSD of the microphone signals can be rewritten as

$$\mathbf{\Phi}_{\mathbf{XX}}(k,m) \approx \Phi_{S_0S_0}(k,m)\,\mathbf{H}(k)\,\mathbf{H}^{\mathbf{H}}(k) + \mathbf{\Phi}_{\mathbf{NN}}(k). \tag{12}$$

In the case of $\Phi_{S_0S_0}(k,m) \neq 0$ the SEVP (10) can be reformulated as follows

$$\mathbf{\Phi}_{\mathbf{NN}}^{-\mathbf{1}}(k)\,\mathbf{H}(k)\,\mathbf{H}^{\mathbf{H}}(k)\,\mathbf{F}(k,m) = \frac{\lambda(k,m)-1}{\Phi_{S_0S_0}(k,m)}\mathbf{F}(k,m). \tag{13}$$

As the rank of the positive semidefinite matrix $\mathbf{\Phi}_{\mathbf{NN}}^{-\mathbf{1}}(k)\,\mathbf{H}(k)\,\mathbf{H}^{\mathbf{H}}(k)$ is one there is obviously only one eigenvector belonging to an eigenvalue greater than zero. This eigenvector which represents the optimal transfer functions $\mathbf{F}_{\text{SNR}}(k)$ is independent of the block number m:

$$\mathbf{F}_{\text{SNR}}(k) = \zeta(k)\,\mathbf{\Phi}_{\mathbf{NN}}^{-\mathbf{1}}(k)\,\mathbf{H}(k), \tag{14}$$

where $\zeta(k)$ is an arbitrary complex constant. This can be easily verified by plugging (14) into (13).

2.2. Generalized Eigenvector Blocking Matrix (GEVBM)

The new GEVBM uses the optimal transfer functions $\mathbf{F}_{\text{SNR}}(k)$ to construct a projection into the orthogonal complement of $\mathbf{H}(k)$. It is, as previously mentioned, intended to produce noise reference signals orthogonal to a speech reference. The DFT of this reference signal is created as

$$Y_{\text{SNR}}(k) := \mathbf{F}_{\text{SNR}}^{H}(k)\mathbf{X}(k). \tag{15}$$

With the projection vector $\mathbf{P}(k)$ the noise references

$$\mathbf{U}(k) := \mathbf{X}(k) - \mathbf{P}(k)Y_{\text{SNR}}(k) \tag{16}$$

should approximately meet the following orthogonality condition for the STDFTs of the noise and speech reference signals

$$E\left[\mathbf{U}(k,m)Y_{\text{SNR}}^{*}(k,m)\right] \overset{!}{\approx} \mathbf{0}, \tag{17}$$

where $(\cdot)^*$ denotes complex conjugation, resulting in

$$\mathbf{P}(k) := \frac{\mathbf{\Phi}_{\mathbf{NN}}(k)\,\mathbf{F}_{\text{SNR}}(k)}{\mathbf{F}_{\text{SNR}}^{\mathbf{H}}(k)\,\mathbf{\Phi}_{\mathbf{NN}}(k)\,\mathbf{F}_{\text{SNR}}(k)}. \tag{18}$$

when using (9). Taking the noise signals (16) as output of the blocking matrix

$$\mathbf{B}^{\mathbf{H}}(k) := \mathbf{I}_M - \mathbf{P}(k)\,\mathbf{F}_{\text{SNR}}^{\mathbf{H}}(k), \tag{19}$$

where $\mathbf{I}_M$ is the identity matrix of dimension M, the novel structure of the blocking matrix is given with (14) and (18) as

$$\mathbf{B}^{\mathbf{H}}(k) := \mathbf{I}_M - \frac{\mathbf{H}(k)\,\mathbf{H}^{\mathbf{H}}(k)\,\mathbf{\Phi}_{\mathbf{NN}}^{-\mathbf{1}}(k)}{\mathbf{H}^{\mathbf{H}}(k)\,\mathbf{\Phi}_{\mathbf{NN}}^{-\mathbf{1}}(k)\,\mathbf{H}(k)}. \tag{20}$$

It can be easily veryfied, that the noise reference signals

$$\begin{aligned}\mathbf{U}(k) =& \mathbf{B}^{\mathbf{H}}(k)\,\mathbf{X}(k)\\ =& \left(\mathbf{I}_M - \frac{\mathbf{H}(k)\,\mathbf{H}^{\mathbf{H}}(k)\,\mathbf{\Phi}_{\mathbf{NN}}^{-\mathbf{1}}(k)}{\mathbf{H}^{\mathbf{H}}(k)\,\mathbf{\Phi}_{\mathbf{NN}}^{-\mathbf{1}}(k)\,\mathbf{H}(k)}\right)\mathbf{N}(k)\end{aligned} \tag{21}$$

do not contain any instationary speech signal components.

Note, because of the indirect estimation of the TFs in (14) high flexibility in constructing a BM is given, e.g. TFRBM [5] could be easily realized.

2.3. Estimation of the cross PSD matrices

For the determination of $\mathbf{F}_{\text{SNR}}(k)$ the PSD matrices appearing in the GEVP (9) are required. The estimation of $\mathbf{\Phi}_{\mathbf{NN}}(k)$ can be performed in periods when only noise is present in the microphone signals. Such periods have to be indicated by a voice activity detection (VAD). One possibility is then to average the instantaneous estimates of K_N frames such as

$$\hat{\mathbf{\Phi}}_{\mathbf{NN}}(k) := \frac{1}{K_N}\sum_{m=1}^{K_N}\left(\mathbf{X}(k,m)\,\mathbf{X}^{\mathbf{H}}(k,m)\right)|_{\mathbf{X}=\mathbf{N}}. \tag{22}$$

Instead of estimating $\mathbf{\Phi}_{\mathbf{XX}}(k,m)$ for each block m it is sufficient to estimate an averaged cross PSD matrix

$$\hat{\mathbf{\Phi}}_{\mathbf{XX}}(k) := \frac{1}{K_X}\sum_{m=1}^{K_X}\mathbf{X}(k,m)\,\mathbf{X}^{\mathbf{H}}(k,m) \tag{23}$$

from the observation of K_X frames in which the microphone signals contain both speech and noise signal parts. The reason for that is that for sufficient large K_X the matrix $\hat{\mathbf{\Phi}}_{\mathbf{XX}}(k)$ assumes the same form as $\mathbf{\Phi}_{\mathbf{XX}}(k,m)$ in (12) except for the constant $\Phi_{S_0S_0}(k,m)$. But obviously, this constant has no effect on the eigenvectors of the GEVP (9).

3. SIMULATION RESULTS

For the experiments a linear array of $M = 5$ microphones with inter-element distance of $d = 0.04\,\mathrm{m}$ placed in a simulated reverberant enclosure of the size (6 m) x (5 m) x (3 m) was used. The noise and speech sources were placed within the enclosure according to Fig. 2 with angles of $\theta_s = 45^\circ$ for the speech and $\theta_n = 20^\circ$ for the directional noise, repectively. We used 10 TIMIT sentences as speech source, each of a length of about 4 seconds. The noise source signal was a recording of lowpass fan noise.

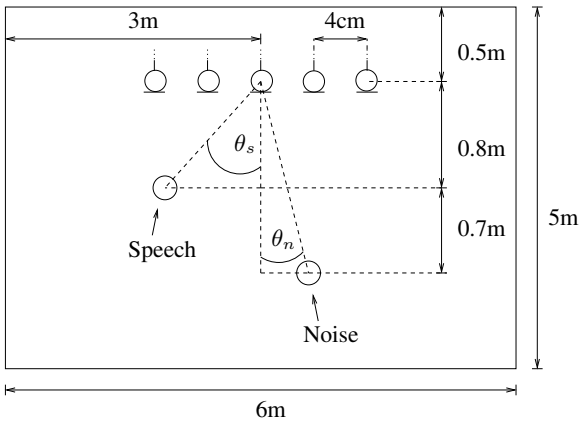

Fig. 2. *Simulation set-up.*

The speech and noise signal components at the microphones were created using the image method by Allen and Berkley and summed with an SNR of 5 dB. Furthermore white gaussian noise was added to each microphone with an SNR of 35 dB. The sampling rate $f = 1/T$ was chosen to 12 kHz.

A delay-and-sum beamformer (DSB) was used as a fixed beamformer. According to the set-up, the delay for the alignment of the speech components was one sampling period T and could therefore be realized perfectly. The multichannel noise cancellation in the AIC was implemented using the normalized least mean squares method with a filter length of 1024 taps.

Four different blocking matrices were compared. For the blocking matrix by Griffith and Jim (GJBM) the microphone signals were aligned the same way as in the DSB. The noise reference signals were computed by subtracting from one microphone signal the mean of the $M - 1$ other aligned microphone signals.

For the estimation of the TFRs in the TFRBM according to [5] the microphone signals were windowed by a Hamming window with a length of $L = 512$ samples and no overlap between successive windows. The estimation time corresponded to the length of each TIMIT sentence. The estimated transfer functions were transformed to the time domain. As the resulting impulse responses were assumed to be noncausal and of finite length, they were cut off to the intervall $[-127, 128]$.

For the GEVBM, first the noise cross PSD matrix was estimated according to (22) in a period of 8 sec. The window length was $L = 512$ samples with an overlap of $L - B = L/2$. The averaged cross PSD matrix of the microphone signals was computed according to (23) when the speech and noise parts were both present. Then the SEVP (10) was solved using the Mises vector iteration applying K_X iterations. The resulting eigenvector was used to form the transfer functions of the GEVBM for each frequency bin as shown in (19). Finally, the FIR filter coefficients were determined the way as in the TFRBM. The filtering in the GEVBM and TFRBM was realized in the frequency domain using the overlap save method.

The adaptive blocking matrix was incorporated into the simulations to get a performance upper bound. In the case of absence of noise during the estimation of the ABM filter coefficients a perfect FIR result is achieved. The ABM [2] was realized in the frequency domain without constraints on the filter coefficients for improved robustness using filter lengths of 256 samples.

For the comparison of the blocking matrices all components of the GSC were analyzed in steady state. That was nearly achieved by choosing a very high number of iterations for the AIC and the ABM. Different measures concerning the noise reduction and speech quality are presented in the following. First of all the SNR gain from the input to the output of the whole GSC structure was computed as a function of the reverberation time T_{60}.

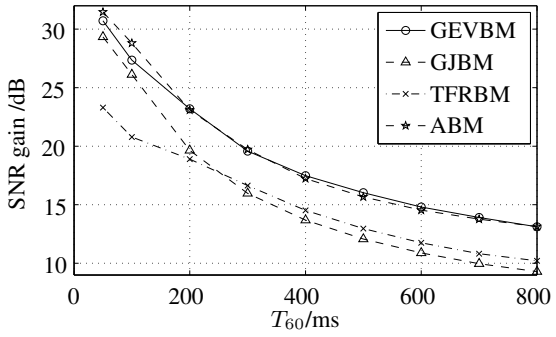

Fig. 3. *SNR gain for directional noise.*

As can be seen in Fig. 3 the the best SNR improvents for all reverberation times were achieved with the ABM. While the SNR gains for the GEVBM are comparable to those of the ABM, the SNR gains of the GJBM are significantly lower compared to ABM and GEVBM for increased reverberation times. For small values of T_{60}, the SNR gains are nearly the same, because the assumption of a simple direct path propagation in the GJBM is then only slightly violated. It is important to mention that the GJBM would loose in performance if, instead of exactly one sampling period, a fractional delay had to be realized by interpolation FIR filters, and if the estimation of the DoA were not perfect.

Unfortunately the SNR gains achieved with the TFRBM are considerably smaller for low reverberation times compared to the other approaches because of less noise reduction in the lower frequencies. For large values of T_{60} the SNR improvements are only slightly higher than those obtained with the GJBM.

One possibility to assess speech signal distortions is to measure the power spectral density deviation between the speech signal components at the beamformer output and a reference signal. Here, we take the ABM-GSC output as the reference as we assume it to produce the best results that can be achieved using FIR filters of a given length. For the k-th frequency bin the PSD devation is defined by

$$\overline{\delta}_{\mathrm{PSD}}(k) = \frac{\overline{\Phi}_{SS}(k)}{\overline{\Phi}_{SS}^{(\mathrm{ABM})}(k)}, \tag{24}$$

where

$$\overline{\Phi}_{SS}(k) := \frac{1}{K_S} \sum_{m=1}^{K_S} |Y_s(k,m)|^2 \tag{25}$$

is the average over the instantaneous estimates of the PSD of the speech signal components $y_s(l)$ of the GSC output signal. K_S denotes the number of windows used for the average. For example,

Fig. 4 displays the PSD deviation over frequency for the reverberation time $T_{60} = 100\,\text{ms}$. It is noticeable that for the TFRBM lower

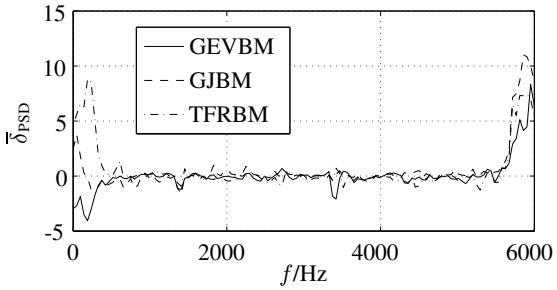

Fig. 4. *Mean PSD deviation of compared blocking matrices for directional noise ($T_{60} = 100\,ms$).*

frequencies under 500 Hz were highly amplified. This fact was already mentioned in [7]. While the amplification of those frequencies was smaller with the GJBM, the GEVBM caused an attenuation of those frequencies. However, the attenuation was smaller than the amplification. Furtheron, an amplification of frequencies over 5 kHz could be observed for all blocking matrix methods.

Taking into account the fact that speech signals have the most power between 500 Hz and 5 kHz the variance of the PSD deviation $\overline{\sigma}_{\text{PSD}}$ averaged over all frequencies in this interval was analyzed, see Fig. 5. A variance of zero means that all frequencies are attenuated or amplified the same way, causing a simple amplitude change of the speech signal. A high value for the variance means that different frequencies are amplified by different factors resulting in speech distortions. Here, it can be seen that the variance for the GJBM

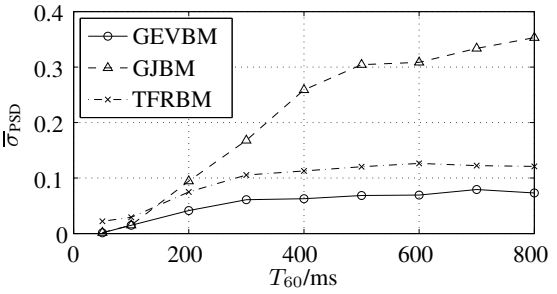

Fig. 5. *Mean PSD deviation variance for directional noise.*

increases significantly with reverberation time, while the TFRBM achieves considerably lower values, and the GEVBM even lower vales than the TFRBM, indicating less speech distortion.

Finally the speech distortion in the GSC output signal is measured by a perceptual speech quality measure (PSM) [8] in Fig. 6. PSM has been shown to give comparable objective perceptual quality evaluation results as the well-known PESQ measure. Here, the reference was the clean speech output signal of the GSC with optimal blocking matrix (ABM). The PSM values for the novel GEVBM and the GJBM are similarly high for low reverberation times. However, for increasing reverberation times the PSM values for GEVBM are slightly higher than for GJBM. The TFRBM delivers somewhat inferior results, which can be explained by the fact that the TFRBM boosts low-frequency signal components. The displayed PSM results correspond well to our informal listening tests.

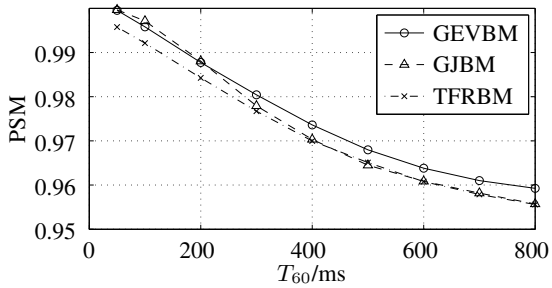

Fig. 6. *Perceptual speech quality measure (PSM) of the GSC output signals using different blocking matrices in the case of directional noise.*

The C/C++ implementation of the GEVBM-GSC for realtime application in our laboratory comprised a VAD, DoA estimation by eigenvalue decomposition, the DSB as FB and the multichannel audio input/output management. The computational effort for the whole system running on an Intel Quad-Core Xeon E5345 / 2.33 GHz processor resulted in a realtime factor of 0.3.

4. CONCLUSION

In this paper a new blocking matrix for a GSC beamformer was presented, which is based on a generalized eigenvalue decomposition. The simulation results show that, compared to other methods, a higher noise reduction and smaller desired signal distortion in the interesting frequency band can be achieved. The performance is similar to that of the adaptive blocking matrix by Hoshuyama et al., while, unlike the latter, adaptation can be carried out in the presence of stationary noise. However, these improvements can only be achieved with a higher computational effort.

5. REFERENCES

[1] L. J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming", *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27-34, Jan. 1982.

[2] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters", *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2677-2684, Oct. 1999.

[3] W. Herbordt and W. Kellermann, "Efficient frequency-domain realization of robust generalized sidelobe cancellers", *Proc. Int'l Workshop on Acoustic Echo and Noise Control IWAENC*, Sep. 2001.

[4] W. Herbordt and W. Kellermann, "Analysis of blocking matrices for generalized sidelobe cancellers for non-stationary broadband signals", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing ICASSP-02*, May. 2002.

[5] S. Gannot, D. Burshtein and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech", *IEEE Trans. on Signal Proc.*, vol. 49, no. 8, pp. 1614-1626, Aug. 2001.

[6] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition", *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 15, no. 5, pp. 1529-1539, July 2007.

[7] S. Gannot, D. Burshtein and E. Weinstein, "Analysis of the power spectral deviation of the general transfer function GSC", *IEEE Trans. on Signal Proc.*, vol. 52, no. 4, pp. 1115-1121, Apr. 2004.

[8] R. Huber, "Objective assessment of audio quality using an auditory processing model", Ph.D. thesis, University of Oldenburg, Oldenburg, 2003.