CONTROLLING SPEECH DISTORTION IN ADAPTIVE FREQUENCY-DOMAIN PRINCIPAL EIGENVECTOR BEAMFORMING

Ernst Warsitz and Reinhold Haeb-Umbach

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

ABSTRACT

Broadband adaptive beamformers, which use a narrowband SNR-maximization optimization criterion for noise reduction, typically cause distortions of the desired speech signal at the beamformer output. In this paper two methods are investigated to control the speech distortion by comparing the eigenvector beamformer with a maximum likelihood beamformer: One is an analytic solution for the ideal case of absence of reverberation and the other one is a statistically motivated approach. We use the recently introduced gradient-ascent algorithm for adaptive principal eigenvector beamforming and then normalize the filter coefficients by the proposed distortion control methods. Experimental results in terms of the achievable SNR gain and a perceptual speech quality measure are given for the normalized eigenvector beamformer and are compared to standard beamforming methods.

1. INTRODUCTION

Hands-free speech communication is often impaired by acoustical background noise of a reverberant enclosure. Whereas single-channel techniques can only utilize spectral information, multi-channel speech enhancement by acoustical beamforming exploits the spectral and spatial diversity to discriminate between desired and undesired signal components. Statistically optimum beamformers like minimum mean square error (MMSE) beamformers, minimum variance distortionless response (MVDR) beamformers or eigenbeamformers are well-known to result in the same weight vector up to a scalar constant [1], which can be realized as a single-channel post-filter. Especially, if the frequency-domain narrowband eigenbeamformer method is applied to broadband speech signals, the resulting filter coefficients correspond to the optimal (in the MVDR sense) filter coefficients up to an unknown scalar per frequency bin. Generally, spatial constraints are designed to assure a distortionless response for the desired signal. Therefore it is necessary to estimate the speaker position or at least the direction-of-arrival (DOA), which is a difficult task on its own right in reverberant environments [2].

An adaptive linearly constraint solution of the MVDR beamformer was originally given by Frost [3] and transformed by Griffith and Jim into an unconstrained efficient realization as generalized sidelobe canceller (GSC) [4]. Here the underlying assumption is that delay-only propagation paths are present between the distant source and the sensors. The reverberation found in practice can then lead to severe signal cancellation. The signal cancellation problem has been addressed in many ways, see [5] for a recent overview. Further, a major limitation of the GSC is the relatively small signal-to-noise ratio improvement in diffuse noise [6]. Unlike the GSC, a data-independent design of the MVDR beamformer for theoretically well-defined sound fields can be done. A study of the practically very relevant case of diffuse backround noise is given in [7]. Because of the ability of the MVDR beamformer to suppress a diffuse noise field it is also known as superdirective beamformer.

Recently we have proposed a Filter-and-Sum beamformer [8] which extracts adaptively the principal eigenvector incorporating the cross power spectral density matrices of speech-plus-noise signal and the noise-only signal at the microphones. The adaptation works blindly, i.e. no explicit source localization is required and the exact microphone positions need not be known. In [9] we have shown that in addition to the direct path, also early reflections are aligned. Furthermore, significant signal-to-noise ratio improvements have been achieved even for diffuse noise environments [8, 10].

In this paper we address in particular the problem of speech distortion at the output of the frequency-domain principal generalized eigenvector beamformer (GEV). Different solutions to the problem will be discussed and verified by SNR and objective speech quality measures in the presence of a directional noise field.

2. STATISTICALLY OPTIMUM BEAMFORMER

We are given an array of M microphones. Each frequency-domain microphone signal $X_i(k)$, $i = 1, \ldots, M$, where k denotes the frequency bin, is assumed to consist of two components: a signal component $S_i(k)$ and a stationary noise term $N_i(k)$. The beamformer output $Y(k)$ is then given by

$$Y(k) = \sum_{i=1}^{M} F_i^*(k) \cdot X_i(k) = \sum_{i=1}^{M} F_i^*(k) \cdot (S_i(k) + N_i(k)). \tag{1}$$

Here, $F_i^*(k)$ is the complex conjugate filter coefficient of the i-th microphone signal. The frame index has been omitted for ease of notation. In the following we will use the vector notation, i.e. $\mathbf{X}(k) = (X_1(k), .., X_M(k))^T$, $\mathbf{F}(k) = (F_1(k), .., F_M(k))^T$, such that (1) can be written as:

$$Y(k) = \mathbf{F}^H(k) \cdot \mathbf{X}(k), \tag{2}$$

where $(\cdot)^T$ denotes transpose and $(\cdot)^H$ Hermitian transpose. If the desired signal $S_i(k)$ and the noise $N_i(k)$ are uncorrelated the power spectral density (PSD) of the beamformer output can be written as

$$\begin{aligned} \Phi_{YY}(k) &= \mathbf{F}^H(k)\mathbf{\Phi_{XX}}(k)\mathbf{F}(k) \\ &= \mathbf{F}^H(k)\mathbf{\Phi_{SS}}(k)\mathbf{F}(k) + \mathbf{F}^H(k)\mathbf{\Phi_{NN}}(k)\mathbf{F}(k), \end{aligned} \tag{3}$$

where $\mathbf{\Phi_{XX}}(k)$, $\mathbf{\Phi_{SS}}(k)$ and $\mathbf{\Phi_{NN}}(k)$ are the cross power spectral density matrices of the microphone signals, the speech and noise terms, respectively. Our goal is to determine a vector of filter coefficients $\mathbf{F}(k)$ such that the signal-to-noise ratio

$$\mathrm{SNR}(k) = \frac{\mathbf{F}^H(k)\mathbf{\Phi_{XX}}(k)\mathbf{F}(k)}{\mathbf{F}^H(k)\mathbf{\Phi_{NN}}(k)\mathbf{F}(k)} - 1 \tag{4}$$

of the output signal $Y(k)$ is maximized. Obviously, the frequency dependent SNR (4) is maximized by the eigenvector corresponding to the largest eigenvalue of $\mathbf{\Phi_{NN}^{-1}}(k)\mathbf{\Phi_{XX}}(k)$. Then the Rayleigh quotient in (4) takes the magnitude of exactly the largest eigenvalue.

Let $\Phi_{UU}(k)$ denote the PSD of the source speech signal $U(k)$ and let $H_i(k)$ be the transfer function from the source to the i-th sensor. The M transfer functions are arranged in a vector $\mathbf{H}(k) = (H_1(k), \ldots, H_M(k))^T$. Then the power spectral density matrix of the sensor signals is given by

$$\mathbf{\Phi_{XX}}(k) = \Phi_{UU}(k)\mathbf{H}(k)\mathbf{H}^H(k) + \mathbf{\Phi_{NN}}(k). \tag{5}$$

We find, that

$$\mathbf{F}_{\mathrm{SNR}}(k) = \mathbf{\Phi_{NN}^{-1}}(k)\mathbf{H}(k) \tag{6}$$

is the principal eigenvector of $\mathbf{\Phi_{NN}^{-1}}(k)\mathbf{\Phi_{XX}}(k)$, which maximizes the frequency dependent beamformer output SNR.

In an alternative design, if the target speaker direction θ_t is known the array steering vector

$$\mathbf{d}(\theta_t, k) = (1,\ e^{-j\omega_k\tau_2(\theta_t)}, \ldots, e^{-j\omega_k\tau_M(\theta_t)})^T \tag{7}$$

can be used as a spatial constraint to ensure a distortionless response for signals arriving from the DOA θ_t: $\mathbf{F}_{MVDR}^H\mathbf{d}(\theta_t, k) = 1$. Here, ω_k is the discrete frequency variable and $\tau_i(\theta_t)$ denotes the delay of the target signal for the i-th sensor relative to the signal at the first sensor, which is a function of the DOA θ_t. With the spatial constraint the so-called minimum variance distortionless response (MVDR) filter vector

$$\mathbf{F}_{\mathrm{MVDR}}(k) = \frac{\mathbf{\Phi_{NN}^{-1}}(k)\mathbf{d}(\theta_t, k)}{\mathbf{d}^H(\theta_t, k)\mathbf{\Phi_{NN}^{-1}}(k)\mathbf{d}(\theta_t, k)} \tag{8}$$

can be derived [3]. The MVDR beamformer ensures a distortionless response for signals arriving from the DOA θ_t and can be shown to be optimum in the Maximum Likelihood sense. Note, that for realistic applications not only a mismatch of the steering direction of the array but also of the sensor position and the sensor gain might cause strong distortions of the desired speech signal at the beamformer output. Robustness to mismatched direction estimates has to be included by widening the spatial selectivity by derivative constraints and/or by diagonal loading of the power spectral density matrices. Hence, the interference rejection will also be reduced.

Comparing (8) with (6) it is seen that, if we set $\mathbf{H}(k) = \mathbf{d}(\theta_t, k)$, i.e. assume a transfer function from source to sensors governed by pure delays, the two beamformers differ only in a scalar constant $w(k)$ per frequency bin k:

$$\mathbf{F}_{\mathrm{MVDR}}(k) = w(k)\mathbf{F}_{\mathrm{SNR}}(k) \tag{9}$$

which is actually a well-known fact [1].

3. FILTER NORMALIZATION

For a distortionless speech signal at the beamformer output we have to find a post-filter $w(k)$ which "normalizes" the filter coefficients $F_{\mathrm{SNR}}(k)$ such that the spatial response of the beamformer $r(\theta, k) = |w(k)\mathbf{F}_{\mathrm{SNR}}^H(k)\mathbf{d}(\theta, k)|$ will have unity gain for any frequency index k in the target direction θ_t: $r(\theta_t, k) \stackrel{!}{=} 1$. We will now derive three alternatives how to determine $w(k)$ and verify the results for the case of no reverberation.

Spatial Normalization (SN) If we assume to know the speaker direction a spatial constraint can be incorporated by normalizing the filter coefficients in (6) by

$$w_{\mathrm{SN}}(k) = \frac{1}{\mathbf{F}_{\mathrm{SNR}}^H(k)\mathbf{d}(\theta_t, k)}. \tag{10}$$

In the case of no reverberation the resulting filter coefficients are

$$w_{\mathrm{SN}}(k)\mathbf{F}_{\mathrm{SNR}}(k)\Big|_{\mathbf{H}(k)=\mathbf{d}(\theta_t,k)} = \frac{\mathbf{\Phi_{NN}^{-1}}(k)\mathbf{d}(\theta_t, k)}{\mathbf{d}^H(\theta_t, k)\mathbf{\Phi_{NN}^{-1}}(k)\mathbf{d}(\theta_t, k)}, \tag{11}$$

which is equal to the MVDR solution given in (8).

Blind Analytical Normalization (BAN) Since we are interested in a blind scheme we now assume that the DOA is no longer known. We propose the following post-filter

$$w_{\mathrm{BAN}}(k) = \frac{1}{\mathbf{F}_{\mathrm{SNR}}^H(k)\mathbf{\Phi_{NN}}(k)\mathbf{F}_{\mathrm{SNR}}(k)}. \tag{12}$$

Using (6) and $\mathbf{H}(k) = \mathbf{d}(\theta_t, k)$ we obtain

$$w_{\mathrm{BAN}}(k)\Big|_{\mathbf{H}(k)=\mathbf{d}(\theta_t,k)} = \frac{1}{\mathbf{d}^H(\theta_t, k)\mathbf{\Phi_{NN}^{-1}}(k)\mathbf{d}(\theta_t, k)}, \tag{13}$$

which will give a distortionless response for the DOA with the filter coefficients $w_{\mathrm{BAN}}(k)\mathbf{F}_{\mathrm{SNR}}(k)$. We call this method *blind analytical normalization* (BAN) since a closed form expression can be given, in contrast to the method proposed next.

Blind Statistical Normalization (BSN) Here, we derive a statistically motivated normalization scheme, which will be shown in the experimental results to have advantages in comparison to $w_{\mathrm{BAN}}(k)$, especially in a reverberant environment. While complete knowledge of the DOA is not available, we might have partial knowledge, which we express by a probability density function $p(\theta)$. Extending (10) to account for such partial knowledge yields

$$w_{\mathrm{BSN}}(k) = \frac{1}{\int_{\theta=-\pi/2}^{\pi/2} p(\theta)|\mathbf{F}_{\mathrm{SNR}}^H(k)\mathbf{d}(\theta, k)|d\theta}. \tag{14}$$

Note that $p(\theta)$ depends on the frequency, which is not indicated in our notation. Also, $\mathbf{F}_{\mathrm{SNR}}(k)$ depends on the DOA, which is also not made explicit in our notation. It is easily seen that eq. (14) reduces to (10), if the DOA is known, i.e. $p(\theta) = \delta(\theta - \theta_t)$. Indeed, some knowledge about θ_t is implicitly available in the filter coefficients: a correctly operating beamformer will have a small spatial response in the direction of the noise and a large response in the direction of the desired signal. Therefore the spatial response itself can be used as probability density function, if properly normalized:

$$p(\theta) = \frac{|\mathbf{F}_{\mathrm{SNR}}^H(k)\mathbf{d}(\theta, k)|}{\int_{\theta=-\pi/2}^{\pi/2} |\mathbf{F}_{\mathrm{SNR}}^H(k)\mathbf{d}(\theta, k)|d\theta}. \tag{15}$$

With (14) and (15) the post-filter for the *blind statistical normalization* (BSN) can be written as

$$w_{\mathrm{BSN}}(k) = \frac{\int_{\theta=-\pi/2}^{\pi/2} |\mathbf{F}_{\mathrm{SNR}}^H(k)\mathbf{d}(\theta, k)|d\theta}{\int_{\theta=-\pi/2}^{\pi/2} |\mathbf{F}_{\mathrm{SNR}}^H(k)\mathbf{d}(\theta, k)|^2 d\theta}. \tag{16}$$

4. ADAPTIVE EIGENVECTOR TRACKING

The determination of the dominant eigenvector of the generalized eigenvalue problem is equivalent to the following constrained optimization problem

$$\max_{\mathbf{F}^H(k)} \mathbf{F}^H(k)\mathbf{\Phi}_{\mathbf{XX}}(k)\mathbf{F}(k) \tag{17}$$

$$\text{subj. to } \mathbf{F}^H(k)\mathbf{\Phi}_{\mathbf{NN}}(k)\mathbf{F}(k) = C(k), \tag{18}$$

where $C(k) \in I\!R^+$ is an arbitrary positive non-zero real constant. In [8] we have derived an iterative gradient ascent algorithm for solving this optimization problem. Omitting the frequency bin index k for ease of notation and introducing the iteration counter κ, which is also the block index of the segmental signal processing, the following gradient ascent algorithm has been obtained:

$$\begin{aligned}\mathbf{F}_{\kappa+1} = \mathbf{F}_\kappa &+ \frac{C - \mathbf{F}_\kappa^H \hat{\mathbf{\Phi}}_{\mathbf{NN}} \mathbf{F}_\kappa}{2\mathbf{F}_\kappa^H \hat{\mathbf{\Phi}}_{\mathbf{NN}} \hat{\mathbf{\Phi}}_{\mathbf{NN}} \mathbf{F}_\kappa} \hat{\mathbf{\Phi}}_{\mathbf{NN}} \mathbf{F}_\kappa \\ &+ \mu \left[\hat{\mathbf{\Phi}}_{\mathbf{XX}} \mathbf{F}_\kappa - \frac{\mathbf{F}_\kappa^H \hat{\mathbf{\Phi}}^{(XN)} \mathbf{F}_\kappa}{2\mathbf{F}_\kappa^H \hat{\mathbf{\Phi}}_{\mathbf{NN}} \hat{\mathbf{\Phi}}_{\mathbf{NN}} \mathbf{F}_\kappa} \hat{\mathbf{\Phi}}_{\mathbf{NN}} \mathbf{F}_\kappa \right],\end{aligned} \tag{19}$$

where $\hat{\mathbf{\Phi}}^{(XN)} = \hat{\mathbf{\Phi}}_{\mathbf{XX}} \hat{\mathbf{\Phi}}_{\mathbf{NN}} + \hat{\mathbf{\Phi}}_{\mathbf{NN}} \hat{\mathbf{\Phi}}_{\mathbf{XX}}$, and μ is the step size parameter. The estimated power spectral density matrices are denoted with $\hat{\mathbf{\Phi}}$ respectively. We assume the noise to be stationary or, at least, to change its statistics on a much larger time scale than the speech signal. Therefore the noise-PSD can be estimated in speech pauses and be still considered a good estimate during speech, whereas the PSD matrix of the reverberated speech has to be estimated during speech periods.

5. EXPERIMENTAL RESULTS

In this section we experimentally evaluate the proposed normalization algorithms of the eigenvector beamformer for speech enhancement in a reverberant enclosure of the size (6 m) x (5 m) x (3 m) in the presence of one additive stationary noise source (recording of computer fan-noise). 10 utterances from different speakers (5 male and 5 female) were used as target speech signals. The sensor signals of the $M = 5$-element linear microphone array were obtained by convolution of speech and noise with simulated room impulse responses for reverberation times T_{60} of 0 s to 0.8 s. The distance between the microphones was 4 cm, and the sampling rate was 12 kHz. The speech source was placed at $\theta_t = 45^\circ$ relative to broadside at a distance of 0.8 m and the noise source at $\theta_n = -20^\circ$ at a distance of 1.6 m. Speech and noise were mixed with a signal-to-noise ratio of about 0 dB. The FIR filters had a length of 128 taps each, and the DFT length was set to 256 taps. Diagonal loading was used for regularization of $\mathbf{\Phi}_{\mathbf{NN}}(k)$ and the integration in (16) was replaced by a sum over 90 discrete angle values.

Filter Normalization First we study the spatial transfer function for the case of no reverberation, known power spectral density matrices and converged filter coefficients. The resulting beampattern obtained by the MVDR beamformer, the eigenvector beamformer without normalization (GEV), with blind analytical normalization (GEV_BAN) and blind statistical normalization (GEV_BSN) are shown in Fig. 1 for different frequencies. It can be seen that a minimum has been put in the direction of the noise and the direction of the main lobe depends on the frequency. Only the MVDR beamformer gives a perfectly distortionless response for the target direction $\theta_t = 45^\circ$.

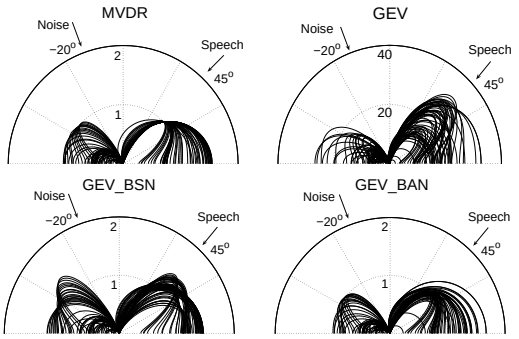

Figure 1: *Beampattern of the MVDR beamformer and the eigenvector beamformer without normalization and with BAN/BSN. Note the different scaling for GEV.*

While the beampattern gives an impression of the spatial response over all angles $\theta = [-\pi/2, .., \pi/2]$, in Fig. 2 the spatial response $r(\theta_t, k) = |\mathbf{F}^H(k)\mathbf{d}(\theta_t, k)|$ is plotted over the frequency: in the upper figure for no reverberation, $T_{60} = 0$ s, and in the lower figure for $T_{60} = 0.3$ s. It can be seen, that the spatial response in the case of no normalization depends highly on the frequency and the reverberation. The BSN post-processing gives good results for all frequencies and reverberation times. The analytical normalization method BAN works well only for low reverberation times.

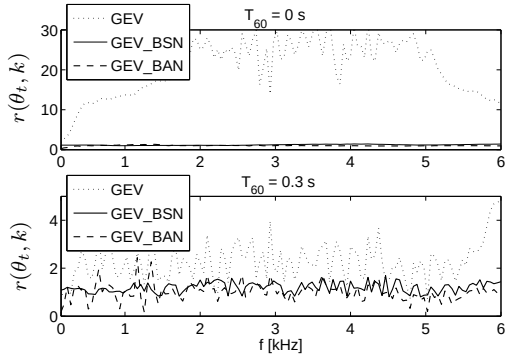

Figure 2: *Spatial response $r(\theta_t, k)$ for the target direction over continuous frequency.*

In Fig. 3 we study the behavior during adaption. It shows the spatial response for θ_t for the frequencies 1, 2, 3, 4 and 5 kHz over time: in the upper figure for no reverberation, $T_{60} = 0$ s, and in the lower figure for $T_{60} = 0.3$ s. While the GEV_BAN beamformer shows high fluctuations during acquisition time the GEV_BSN beamformer exhibits no peaks in the spatial response for all times.

Signal-to-Noise Ratio The SNR gain from the multi-channel beamformer input to the beamformer output for the already introduced beamformer methods (MVDR, GEV, GEV_BAN, GEV_BSN) and a Delay-and-Sum beamformer (DSB) is shown in Fig. 4 over the reverberation time. While a perfect time alignment of the DSB has been done, no compensation of possible level differences of the input channels was made. In the figure the not normalized GEV beamformer gives the highest SNR gain. This is due to the fact, that it boosts the frequencies with high speech power and thus results in a better SNR compared to

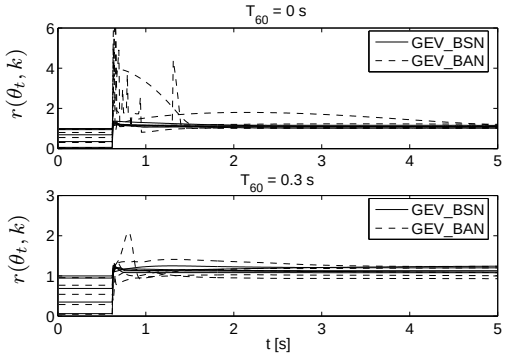

Figure 3: *Spatial response $r(\theta_t, k)$ for the target direction for the frequencies 1, 2, 3, 4 and 5 kHz over time.*

the other methods. Not seen in Fig. 4 is the fact that the SNR gain obtained by GEV_BAN exhibits a strong fluctuation from sentence to sentence for large T_{60}. This makes GEV_BAN unsuitable for higher reverberation times.

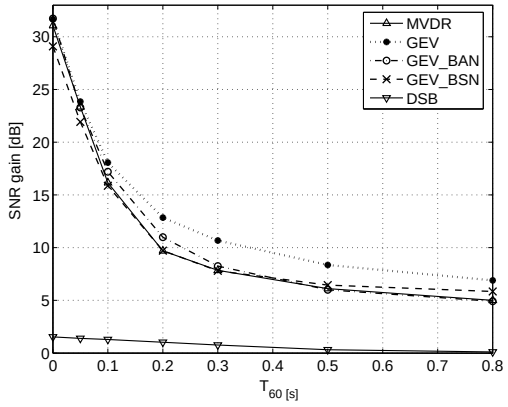

Figure 4: *Comparison of the SNR gain.*

Perceptual Speech Quality Since the SNR gain does not reflect speech distortions, the different beamforming methods are finally compared by the perceptual similarity measure PSM [11] in Fig. 5. PSM has been shown to give comparable objective perceptual quality evaluation results as the well-known PESQ measure [12]. The MVDR beamformer gives for low reverberation times the highest PSM values, although the SNR gain is smaller than for the GEV methods, where the better SNR is bought at the expense of additional speech distortion. With increasing reverberation time the speech distortion decreases and remaining little amplification of some spectral components of the speech results for the GEV beamformer with and without normalization in a better perceptual quality compared to the MVDR.

6. CONCLUSIONS

In this paper we extended the recently proposed generalized eigenvector beamformer by single-channel normalization postprocessing to control the speech distortions. Two different approaches have been realized: one analytic solution for the ideal case of no reverberation and one statistically motivated method. While the GEV beamformer alone gives the highest SNR improvement, also some speech distortion is incorporated. Scaling of the filter coefficients by the proposed normalization methods results in reduction of speech distortion, shown by the perceptual

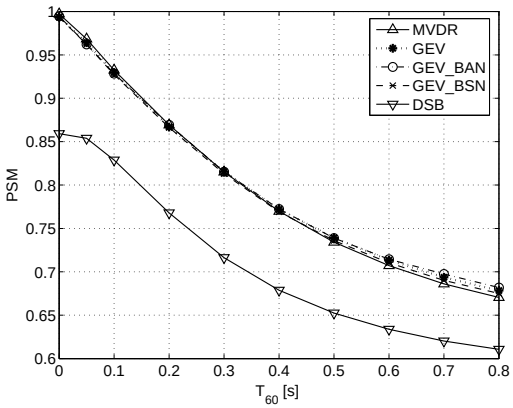

Figure 5: *Comparison of the perceptual similarity measure.*

similarity measure PSM. The advantage of the GEV beamformer is blind adaption, i.e. no explicit estimation of the direction of the desired source and no calibration is needed. Further, the proposed normalization schemes can be used for any adaptive generalized principal eigenvector tracking algorithm, not only the specific method used here.

7. REFERENCES

[1] L.C. Godara, "Applications of Antenna Arrays to Mobile Communications, Part II: Beam-Forming and Direction-of-Arrival Considerations", *Proceedings of the IEEE*, vol. 85, no. 8, pp 1195-1245, Aug. 1997.

[2] J.H. DiBiase, H.F. Siverman and M.S. Brandstein, "Robust localization in reverberant rooms", in *Microphone Arrays: Signal Processing Techniques and Applications*, Springer Verlag, 2001.

[3] O. L. Frost, "An Algorithm for Linearly Constrained Adaptive Array Processing", *Proceedings of the IEEE*, vol. 60, no. 8, pp 926-935, Aug. 1972.

[4] 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.

[5] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition", in *Adaptive Signal Processing – Applications to Real-World Problems*, J. Bensty, Y. Huang (Eds.), Springer, 2003.

[6] J. Bitzer, K.U. Simmer, and K.D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement", in *Proc. IEEE ICASSP*, Phoenix, May 1999.

[7] J. Bitzer and K. U. Simmer, "Superdirective Microphone Arrays", in *Microphone Arrays: Signal Processing Techniques and Applications*, Springer Verlag, 2001.

[8] R. Haeb-Umbach and E. Warsitz, "Adaptive Filter-and-Sum Beamforming in Spatially Correlated Noise", in *Proc. IWAENC*, Eindhoven, Netherlands, Sep. 2005.

[9] E. Warsitz and R. Haeb-Umbach, "Acoustic Filter-and-Sum Beamforming by Adaptive Principal Component Analysis", in *Proc. ICASSP*, Philadelphia, USA, Mar. 2005.

[10] E. Warsitz and R. Haeb-Umbach, "Mehrkanalige Sprachsignalverarbeitung durch adaptives Eigenbeamforming fuer Freisprecheinrichtungen im Kraftfahrzeug", in *Proc. DAGA*, Braunschweig, March, 2006

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

[12] T. Rohdenburg, V. Hohmann and B. Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes", in *Proc. IWAENC*, Eindhoven, Netherlands, Sep. 2005.