OUTLIER-ROBUST DFT-DOMAIN ADAPTIVE FILTERING FOR BIN-WISE STEPSIZE CONTROLS, AND ITS APPLICATION TO A GENERALIZED SIDELOBE CANCELLER

W. Herbordt[1], *H. Buchner*[2], *S. Nakamura*[1], *and W. Kellermann*[2]

[1]ATR Spoken Language Translation Research Laboratories, Kyoto, Japan
{wolfgang.herbordt, satoshi.nakamura}@atr.jp

[2]Telecommunications Laboratory, University Erlangen-Nuremberg, Germany
{buchner,wk}@LNT.de

ABSTRACT

The sensitivity against outliers due to undetected noise bursts often limits the performance of adaptive filters in applications where especially fast convergence is required, as, for example, in adaptive beamforming for audio signal acquisition or in acoustic echo cancellation. In this paper, we analyze the problem of outliers for a robust generalized sidelobe canceler (RGSC) in the discrete Fourier transform (DFT) domain and show by experimental results how outlier-robust adaptive filtering for bin-wise double-talk detectors improves the performance of the RGSC.

1. INTRODUCTION

The choice of an adaptive filter for applications such as acoustic echo cancellation or adaptive beamforming is mostly determined by the convergence speed, the tracking capability, the computational complexity, and the delay. Additionally, for echo cancellation, in [1], robustness against undetected double-talk bursts due to the presence of local sources was pointed out and addressed by using a non-linear function of the error signal for the adaptation. In [2, 3, 4, 5, 6], robustness against undetected double-talk ('outliers') is obtained by deriving adaptive filters from outlier-robust optimization criteria [7].

In recent years, discrete Fourier transform (DFT) domain adaptive algorithms ('frequency-domain adaptive filters' (FDAFs)) have become very attractive since they combine fast convergence with low computational complexity [8]. DFT-domain realizations of acoustic echo cancellers also allow for a DFT-bin-wise adaptation. This is especially advantageous for signals which are sparse in the time-frequency domain, since the stepsize of the adaptive algorithm can be adjusted for each DFT-bin individually. This leads to a more frequent adaptation and faster convergence of the adaptive filter [9]. To improve the robustness of this class of algorithm, a robust DFT-domain adaptive filter based on robust statistics and a non-linear least-squares error (LSE) criterion is derived and applied to acoustic echo cancellation in [5]. However, due to the time-domain optimization criterion, [5] cannot be used in combination with a DFT-bin-wise stepsize control. Therefore, in [10], an outlier-robust DFT-domain adaptive filter for multi-channel systems ('multi-channel bin-wise robust FDAF', MC-BRFDAF) is derived based on a cost function in the DFT domain so that DFT bin-wise stepsize controls can be used for outlier-robust algorithms. The efficiency of the approach was verified by applying the MC-BRFDAF to adaptive beamforming for multi-channel speech enhancement with microphone arrays using a DFT-domain robust generalized sidelobe canceller (RGSC) [11].

In this paper, we statistically analyze the outlier problem of the RGSC and illustrate the performance improvement of the RGSC which is obtained by using MC-BRFDAF instead of MC-FDAF. In Sect. 2, we describe the motivation for using outlier-robust adaptive filters for the RGSC by statistically analyzing the adaptation control of the RGSC. In Sect. 3, we summarize the derivation of MC-BRFDAF starting with outlier-robust maximum likelihood estimation. In Sect. 4, the MC-BRFDAF is applied to the RGSC and experimental results are reported.

We use the following conventions: Upper case and lower case bold font denote matrix and vector quantities, respectively. Underlined quantities denote DFT-domain variables, N is the number of filter taps of an adaptive filter. n is the frequency index, and $2N$ is the DFT size. k is the discrete time index, and r is the block time index. R is the block shift in samples. r is related to k by $k = rR$.

2. ROBUST GENERALIZED SIDELOBE CANCELLER (RGSC) WITH DFT-BIN-WISE DOUBLE-TALK DETECTION

The RGSC, consisting of a fixed beamformer, the adaptive blocking matrix (BM) $\underline{\mathbf{B}}(r)$ after [12] and an interference canceller (IC) $\underline{\mathbf{a}}(r)$ is depicted in Fig. 1. As pointed out in [12], the blocking matrix should be adapted when only the desired signal is present, while the interference canceller should be adapted when only interference is present to prevent instability of the adaptive filters and cancellation of the desired signal. For optimally tracking the time-variance of the sensor signals, the sparseness of the sensor signals in the DFT domain may be exploited, which requires (1) a DFT bin-wise classifier for 'desired signal only', 'interference only', and 'double-talk' between the desired signal and interference, and (2) DFT-domain adaptive filters for adapting $\underline{\mathbf{B}}(r)$ and $\underline{\mathbf{a}}(r)$.

2.1. Bin-Wise Double-Talk Detection

A DFT bin-wise classifier for 'desired signal only', 'interference only', and 'double-talk' is presented in [13]. The classifier exploits the directivity of a fixed beamformer which is steered to the position of the desired source. It is thus assumed that the position of the desired source is roughly known, as, for example, in scenarios where the microphone array is mounted on a computer screen.

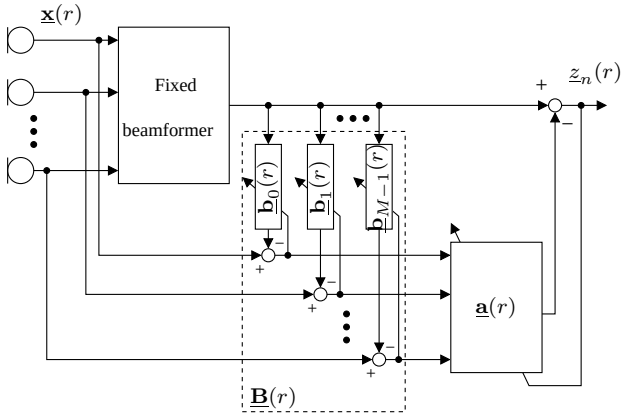

Figure 1: RGSC with an adaptive blocking matrix after [12].

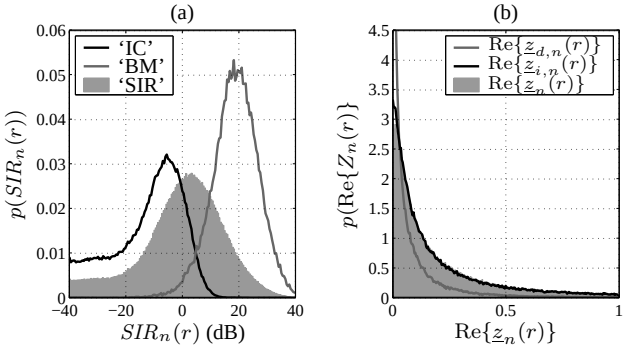

Figure 2: (a) Histogram of $SIR_n(r)$ averaged over frequency n and over block time r: 'SIR' for all samples, 'BM' for samples for which the blocking matrix is adapted, and 'IC' for samples for which the interference canceller is adapted. (b) Histogram of the real part of $\underline{z}_n(r) = \underline{z}_{d,n}(r) + \underline{z}_{i,n}(r)$, of $\underline{z}_{d,n}(r)$, and of $\underline{z}_{i,n}(r)$ for samples for which the interference canceller is adapted; Speech signals for the desired signal $\underline{z}_{d,n}(r)$ and for the interference $\underline{z}_{i,n}(r)$.

A statistical analysis of this classifier is depicted in Fig. 2. In Fig. 2a, the histograms of the signal-to-interference ratio (SIR) for which the classifier detects 'desired signal only' (gray line) and 'interference only' (black line) along with the histogram of the SIR at the sensors (gray shaded surface) are shown. The histograms are averaged over DFT bins n and block-time r for two competing speech signals of length 20 s with average $SIR = 3$ dB. It may be seen that – although the centers of the histograms for 'desired signal only' and 'interference only' are located at high $SIR_n(r) = 20$ dB and at low $SIR_n(r) = -5$ dB, respectively – the variances of the histograms cannot be neglected and 'desired signal only' and 'interference only' are wrongly detected for low $SIR_n(r)$ and high $SIR_n(r)$, respectively.

These wrong detections lead to the adaptation of the blocking matrix and of the interference canceller during the presence of interference and desired signal, respectively, which should be avoided to prevent instabilities and cancellation of the desired signal.

As an example, in Fig. 2b, the distribution of the outliers for the adaptation of the interference canceller is illustrated: The histogram (gray shading) of the real part of the DFT coefficients $\underline{z}_n(r) = \underline{z}_{d,n}(r) + \underline{z}_{i,n}(r)$ of the RGSC output, where $\underline{z}_{d,n}(r)$ and $\underline{z}_{i,n}(r)$ are the DFT coefficients of the desired signal and of the interference, respectively, are depicted for the samples for which the interference canceller is adapted. The histograms of $\mathrm{Re}\{\underline{z}_{d,n}(r)\}$ (gray line) and of $\mathrm{Re}\{\underline{z}_{i,n}(r)\}$ (black line) correspond to the outliers and to the undisturbed data, respectively.

When adapting $\underline{\mathbf{B}}(r)$ and $\underline{\mathbf{a}}(r)$ by a conventional non-robust adaptation algorithm, instability is avoided by using small adaptation stepsizes and accepting in turn a slower convergence of the adaptive filters. To prevent instability while assuring fast convergence of the adaptive filters, we apply outlier-robust adaptive filters which are based on outlier-robust maximum likelihood (ML) estimation [7].

3. DOUBLE-TALK RESILIENT DFT-DOMAIN ADAPTIVE FILTER

In outlier-robust ML estimation, the outliers are assumed to have a distribution D which belongs to some given parametric family $\mathcal{D}$. The estimator is designed to obtain the best signal reconstruction for the least favorable outlier distribution within the given parametric family. Huber defines an ϵ-contaminated normally distributed data set $\mathcal{P}_\epsilon$ as [7]

$$\mathcal{P}_\epsilon = \{(1-\epsilon)\Phi + \epsilon D : D \in \mathcal{D}\} \,, \tag{1}$$

where Φ is the normal distribution with zero mean and unity variance, $\mathcal{D}$ is the set of all distributions symmetric to the origin, and $\epsilon \in [0,1]$ is the outlier probability.

Note that two aspects need to be considered when applying the contaminated model $\mathcal{P}_\epsilon$ to DFT coefficients of speech as, for example, in Fig. 2b: (a) The Gaussian distribution Φ only roughly approximates the true distribution of the speech signal $\underline{z}_{i,n}(r)$, which may be modeled more accurately by a Laplacian distribution, and (b) the density of the sum of independent random variables $\underline{z}_n(r) = \underline{z}_{d,n}(r) + \underline{z}_{i,n}(r)$ should rather be modeled by the convolution of densities than by the sum of densities as in (1).

In [7] it is shown that the least favorable distribution in $\mathcal{P}_\epsilon$, in the sense that the asymptotic variance is maximized, is given by

$$p(z) = \frac{(1-\epsilon)}{\sqrt{2\pi}} \exp\{-\rho(|z|)\} \,, \tag{2}$$

$$\rho(|z|) = \begin{cases} \frac{z^2}{2} & \text{for } |z| \le k_0 \,, \\ k_0|z| - \frac{k_0^2}{2} & \text{for } |z| > k_0 \,, \end{cases} \tag{3}$$

where the constant k_0 depends on ϵ and is chosen such that $\int_{-\infty}^{\infty} p_H(z)dz = 1$. It may be seen that the least favorable distribution is Gaussian in the center and Laplacian in the tails. The transition depends on ϵ and decreases with increasing ϵ.

Interpreting z as the error signal $\underline{e}_n(r)$ of an optimum linear filter $\underline{\mathbf{w}}(r)$, an M-estimator (or maximum likelihood type estimator) [7] of $\underline{\mathbf{w}}(r)$ can be derived by minimizing the cost function

$$\xi(r) = \sum_{n=0}^{2N-1} -\log p\left(\frac{|\underline{e}_n(r)|}{s_n(r)}\right) \tag{4}$$

w.r.t. $\underline{\mathbf{w}}(r)$, or, equivalently,

$$\xi(r) = \sum_{n=0}^{2N-1} \rho\left(\frac{|\underline{e}_n(r)|}{s_n(r)}\right) \,. \tag{5}$$

The scale factor $s_n(r)$ normalizes the variance of the argument of $\rho(\cdot)$, as required in (1). For $|\underline{e}_n(r)|/s_n(r) \le k_0$, Eq. (5)

corresponds to an LSE criterion with a quadratic cost function, while (5) is a 1-norm criterion for $|\underline{e}_n(r)|/s_n(r) > k_0$. For $|\underline{e}_n(r)|/s_n(r) > k_0$, which very likely corresponds to outliers, the gradient of $\xi(r)$ is limited so that the robustness against outliers is increased.

Equation (5) can be solved by an iterative Newton algorithm [14] of the form

$$\underline{\mathbf{w}}(r) = \underline{\mathbf{w}}(r-1) - \underline{\boldsymbol{\mu}}(r)\underline{\boldsymbol{\Lambda}}^{-1}(r)\nabla\xi(r)\,. \qquad (6)$$

$\nabla\xi(r) = 2\partial\xi(r)/\partial\underline{\mathbf{w}}^*(r)$ is the gradient of the cost function $\xi(r)$ w.r.t. $\underline{\mathbf{w}}(r)$. $\underline{\boldsymbol{\Lambda}}(r) = \mathcal{E}\{\nabla^2\xi(r)\} = 4\mathcal{E}\{\partial^2\xi(r)/\partial^2\underline{\mathbf{w}}^*(r)\}$ is the expected value of the Hessian of $\xi(r)$ w.r.t. $\underline{\mathbf{w}}(r)$. $\underline{\boldsymbol{\mu}}(r)$ is a diagonal matrix of size $2N \times 2N$ with stepsizes $\mu_n(r)$, $n = 0,\ 1,\ \ldots,\ 2N-1$, on the main diagonal for controling separately the adaptation in the frequency bins. The DFT-domain Newton step (6) is analogous to the Newton step in the discrete time domain in [4] and an extension of the DFT-domain Newton step in [5] to a bin-wise operation. The derivation of the adaptation algorithm (MC-BRFDAF) based on the Newton step (6) can be found in [10].

To estimate the scale parameter $s_n(r)$, we use the outlier-robust M-estimator for scale as presented in [2, 3].

4. EXPERIMENTAL RESULTS

We apply the MC-BRFDAF to the adaptation of the blocking matrix and of the interference canceller of the RGSC and compare the performance with the RGSC using MC-FDAFs. The bin-wise scale parameter $s_n(r)$ is replaced by a bin-independent scale parameter $s(r)$ since the dependency on DFT bins did not improve the performance relative to a bin-independent estimation in our experiments. The fixed beamformer is realized by a simple uniformly-weighted delay & sum beamformer.

4.1. Transient Behavior

In a first experiment, we study the transient behavior of RGSC for a car environment with $T_{60} = 50\,\text{ms}$ with presence of car noise. The transient behavior of the RGSC for an office environment with $T_{60} = 250\,\text{ms}$ and an interfering speech signal is illustrated in [10]. The performance improvement by MC-BRFDAF for interfering speech is greater than for interfering car noise. (See also Sect. 4.2.)

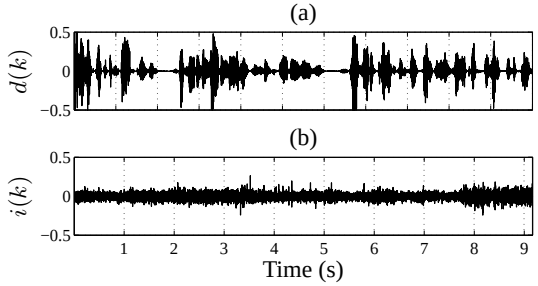

Figure 3: Comparison of the RGSC using MC-FDAF and MC-BRFDAF for 'continuous' double-talk: **(a)** Desired signal $d(k)$ and **(b)** interference $i(k)$ at one microphone.

A microphone array with 12-cm aperture and $M = 4$ microphones is mounted on the sun visor in the passenger cabin of a car. The desired signal (Fig.3a) arrives from the broadside direction from a distance of 60 cm. The car noise is depicted in Fig.3b. The average SIR at the sensors is 6 dB. The frequency range is 200 Hz–6 kHz. The parameters are optimized individually for MC-BRFDAF and for MC-FDAF for maximum convergence speed and maximum noise-suppression after convergence. They are the same for both GSC realizations except for the constant stepsize parameter μ_c, which is used during adaptation of the adaptive filters. (See Fig. 4.)

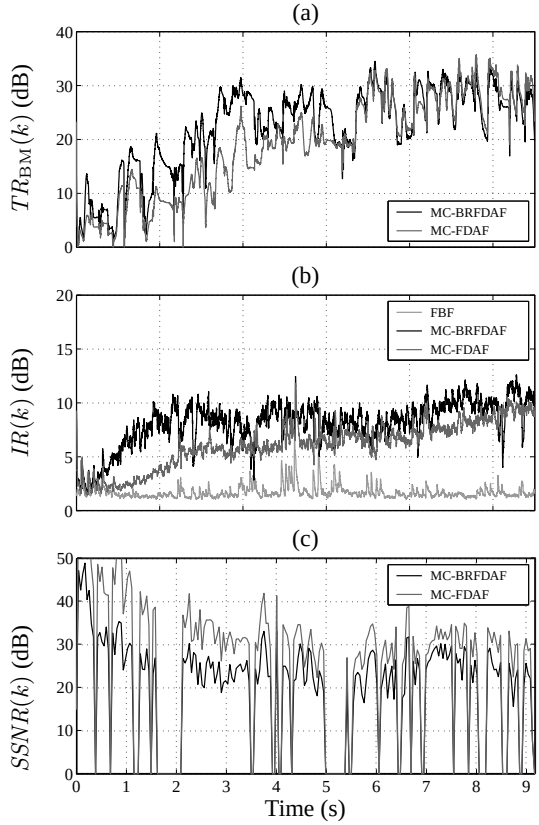

Figure 4: Comparison of the RGSC using MC-FDAF and MC-BRFDAF for 'continuous' double-talk: **(a)** suppression of the desired signal $TR_{\mathrm{BM}}(k)$ by the blocking matrix, **(b)** suppression of the interference $IR(k)$ by the RGSC and by the fixed beamformer (FBF), and **(c)** distortion of the desired signal measured by the segmental SNR $SSNR(k)$ between the FBF output and the RGSC output for data blocks of length 512. (Sampling rate 12 kHz, $N = 256$, $R = 64$, forgetting factor for recursively averaging power spectral densities $\lambda = 0.97$, $k_0 = 1.0$, MC-FDAF: $\mu_c = [0.7, 0.2] \cdot (1-\lambda)$ [BM,IC], MC-BRFDAF: $\mu_c = [1.3, 1.0] \cdot (1-\lambda)$ [BM,IC])

Figures 4a–c show the suppression $TR_{\mathrm{BM}}(k)$ of the desired signal by the blocking matrix, the interference suppression $IR(k)$ of the GSC, and the distortion $SSNR(k)$ of the desired signal by the RGSC as a function of time after initialization of the system, respectively. $SSNR(k)$ is the segmental SNR between the output of the fixed beamformer and the output of the GSC for the desired signal alone with accounting for the delay between the two outputs. Ideally, $SSNR(k) = \infty$ since the interference canceller should not distort the desired signal. It may be seen that the blocking matrix (Fig. 3a) and the interference canceller (Fig. 3b) converge faster for MC-BRFDAF than for MC-FDAF, since a larger stepsize can be chosen for MC-BRFDAF due to the im-

proved robustness against double-talk. In this setup, $TR_{\mathrm{BM}}(k)$ and $IR(k)$ converge for both RGSCs to nearly the same value. One may expect that the larger stepsizes for MC-BRFDAF relative to MC-FDAF may lead to reduced $TR_{\mathrm{BM}}(k)$ and $IR(k)$ after the adaptive filters have reached a steady state. However, due to the dependency of the adaptive filter coefficients on the time-varying spectra of the input signals for this type of interference cancellation problem [11], a larger step size may yield a comparable or even improved steady-state performance because of the faster convergence. The $SSNR(k)$ (Fig. 3c) is lower for MC-BRFDAF than for MC-FDAF. However, for both adaptation algorithms $SSNR(k)$ is large (> 20 dB) so that the distortion is negligible for many applications.

4.2. Steady-State Performance

The interference suppression of the RGSC after convergence of the adaptive filters as a function of the SIR at the sensors is depicted for the office environment with $T_{60} = 250$ ms in Fig. 5a and for the car environment in Fig. 5b. The experimental setup is the same as in Sect. 4.1. In the office, an interfering male speaker is located at 90 degrees off the target. 'Fixed scaling' and 'adaptive scaling' stand for time-invariant scaling $s(r)$ optimized for $SIR = 6$ dB and time-varying scaling using the outlier-robust estimator, respectively. The comparison between 'adaptive scaling' and 'fixed scaling' illustrates the necessity of scale parameter estimation.

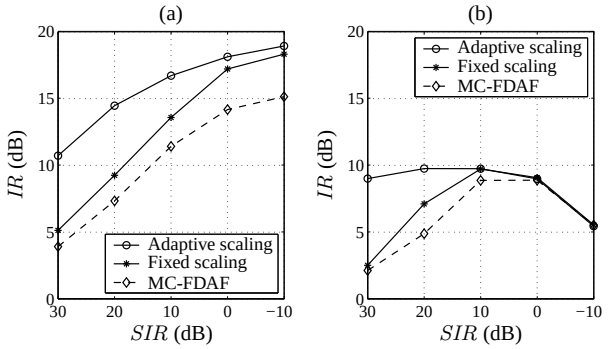

Figure 5: Interference suppression IR after 'convergence' of the IC as a function of the time-averaged SIR for (a) competing speech in an office and (b) car noise in the passenger cabin of a car.

It can be seen that the adaptation of the RGSC by MC-BRFDAF ('adaptive scaling') outperforms MC-FDAF for all SIR. The IR increases with increasing SIR for both environments. For the office environment with competing speech (Fig. 5a), more than 4-dB improvement is obtained for the entire range of SIR. For car noise (Fig. 5b) and low $SIR < 0$ dB, the advantage of MC-BRFDAF relative to MC-FDAF is neglibible. This results from the difficulty of the adaptation control to detect 'interference only' in slowly time-varying diffuse noise fields. The cancellation of the desired signal by the blocking matrix TR_{BM} is not reported for this experiment since, for TR_{BM}, the difference between MC-BRFDAF and MC-FDAF is negligible after convergence of the adaptive filters, as shown by the results in Sect. 4.1.

5. CONCLUSION

We analyzed the problem of outliers for the RGSC and showed how the MC-BRFDAF can be used to resolve this problem. Experimental results show that both the transient behavior and the steady-state performance of the RGSC can be significantly improved by using MC-BRFDAF instead of MC-FDAF.

6. REFERENCES

[1] M.M. Sondhi, "An adaptive echo canceller," *The Bell System Technical Journal*, vol. XLVI, no. 3, pp. 497–510, March 1967.

[2] T. Gänsler, "A double-talk resistant subband echo canceller," *Signal Processing*, vol. 65, no. 1, pp. 89–101, February 1998.

[3] T. Gänsler, S.L. Gay, M.M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 656–663, November 2000.

[4] J. Benesty and T. Gänsler, "A robust fast converging least-squares adaptive algorithm," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 6, pp. 3785–3788, May 2001.

[5] H. Buchner, J. Benesty, T. Gänsler, and W. Kellermann, "An outlier robust extended multidelay filter with application to acoustic echo cancellation," *Int. Workshop on Acoustic Echo and Noise Control*, September 2003.

[6] S. Shimauchi, Y. Haneda, and A. Kataoka, "Frequency-domain adaptive algorithm with non-linear function of error-to-reference ratio for double-talk robust echo cancellation," *Acoustical Science and Technology*, vol. 26, no. 1, pp. 8–15, January 2005.

[7] P.J. Huber, *Robust Statistics*, Wiley, New York, 1981.

[8] J.J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, pp. 14–37, January 1992.

[9] B. H. Nitsch, "A frequency-selective stepfactor control for an adaptive filter algorithm working in the frequency domain," *Signal Processing*, vol. 80, no. 9, pp. 1733–1745, September 2000.

[10] W. Herbordt, H. Buchner, S. Nakamura, and W. Kellermann, "Application of a double-talk resilient DFT-domain adaptive filter for bin-wise stepsize controls to adaptive beamforming," *(submitted to) Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, May 2005.

[11] W. Herbordt, *Sound capture for human/machine interfaces: Practical aspects of microphone array signal processing*, Springer, Heidelberg, Germany, 2005.

[12] 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, October 1999.

[13] W. Herbordt, T. Trini, and W. Kellermann, "Robust spatial estimation of the signal-to-interference ratio for nonstationary mixtures," *Proc. Int. Workshop on Acoustic Echo and Noise Control*, pp. 247–250, September 2003.

[14] S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, Upper Saddle River, NJ, 1993.