Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 2002-163

The impact of speech detection errors on the noise reduction performance of multi-channel Wiener filtering and Generalized Sidelobe Cancellation [1]

Ann Spriet[2], Marc Moonen[3] ,Jan Wouters[4]

Accepted for publication in Signal Processing, September 2004

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.ac.be* in the directory *pub/sista/spriet/reports/02-163.pdf*

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), SISTA, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium, Tel. 32/16/32 18 99, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.ac.be/sista*. E-mail: *ann.spriet@esat.kuleuven.ac.be*. K.U.Leuven, Lab. Exp. ORL/ ENT-Dept., Kapucijnenvoer 33, 3000 Leuven, Belgium, Tel. 32/16/33 24 15, Fax 32/16/33 23 35, WWW: *http://www.kuleuven.ac.be/exporl/Lab/Default.htm*. Ann Spriet is a Research Assistant supported by the Fonds voor Wetenschappelijk Onderzoek (FWO) - Vlaanderen. This research work was carried out at the ESAT laboratory and Lab. Exp. ORL of the Katholieke Universiteit Leuven, in the the frame of IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling'), the Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology)of the Flemish Government, Research Project FWO nr.G.0233.01 ('Signal processing and automatic patient fitting for advanced auditory prostheses') and was partially sponsored by Cochlear. The scientific responsibility is assumed by its authors.

[3]K.U.Leuven, Dept. of Electrical Engineering (ESAT), SISTA, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium, Tel. 32/16/32 17 09, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.ac.be/sista*. E-mail: *marc.moonen@esat.kuleuven.ac.be*. Marc Moonen is a professor at the Katholieke Universiteit Leuven.

[4]K.U.Leuven, Lab. Exp. ORL, Dept. Neurowetenschappen, Kapucijnenvoer 33, 3000 Leuven, Belgium, Tel. 32/16/33 23 42, Fax 32/16/33 23 35, WWW: *http://www.kuleuven.ac.be/exporl/Lab/Default.htm* E-mail: *jan.wouters@uz.kuleuven.ac.be*. Jan Wouters is a professor at the Katholieke Universiteit Leuven.

The impact of speech detection errors on the noise reduction performance of multi-channel Wiener filtering and Generalized Sidelobe Cancellation

A. Spriet [a,b,*], M. Moonen [a] and J. Wouters [b]

[a] *KU Leuven, ESAT/ SCD-SISTA, Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, BELGIUM*

[b] *KU Leuven, ENT-Dept./ Lab. Exp. ORL, Kapucijnenvoer 33, B-3000 Leuven, BELGIUM*

Abstract

In contrast to the Generalized Sidelobe Canceller (GSC), the noise reduction performance of a recently developed multi-channel Wiener filter (MWF) technique does not depend on the validity of a priori assumptions about the signal model. This provides a potential benefit of the MWF over the GSC. However, both techniques also rely on a speech detection algorithm. In this paper, we analyze the average effect of speech detection errors on the performance of the GSC and the MWF both theoretically and experimentally. In the GSC case, it is the simultaneous presence of signal model errors and speech detection errors that affects performance. Incorporating a constraint on the noise sensitivity of the GSC limits the drastic impact of speech detection errors at the expense of reduced noise reduction performance. It is shown that the MWF preserves its benefit over the GSC for a reasonable speech detection error rate of 20% or less, even when the GSC is supplied with a noise sensitivity constraint. Real data experiments confirm the theoretical results.

Key words:
Multi-microphone noise reduction, speech detection, multi-channel Wiener filter, Generalized Sidelobe Canceller
PACS: 43.60.Gk, 43.72.Ew, 43.72Kb

* Corresponding Author. Tel.: +32-16-321795; Fax: +32-16-321970
Email addresses: ann.spriet@esat.kuleuven.ac.be (A. Spriet), marc.moonen@esat.kuleuven.ac.be (M. Moonen), jan.wouters@uz.kuleuven.ac.be (J. Wouters).

1 Introduction

In speech communication applications, such as teleconferencing, handsfree telephony and hearing aids, the presence of background noise and/or reverberation seriously reduces the intelligibility of the desired speech signal. This stems from the large distance between the speaker and the microphone(s). Hence, the use of a noise reduction algorithm is necessary. Multi-microphone systems exploit spatial information in addition to temporal and spectral information of the desired signal and noise signal and are thus preferred to single microphone procedures (such as spectral subtraction). A widely used adaptive multi-microphone noise reduction technique is the Generalized Sidelobe Canceller (GSC) [1–3]. The GSC assumes the desired speaker location, as well as the microphone characteristics and positions to be known a priori, and reflections of the speech signal to be absent. In reality, these conditions are seldom completely fulfilled so that leakage of speech into the noise references occurs, especially at high input Signal-to-Noise Ratios (SNR) [4,5]. To avoid cancellation of speech, the Adaptive Noise Canceller (ANC) in the GSC is adapted during periods of noise only [6,5,7], hence the need for robust speech/noise detection. When used in combination with small-sized arrays, e.g., in hearing aid applications, an additional constraint on the noise sensitivity of the GSC [7–10] is often included to guarantee performance in the presence of small errors in the assumed signal model, e.g. due to microphone mismatch [11]. However, this constraint goes at the expense of less noise reduction.

In [12–16], a *multi-channel Wiener filtering* (MWF) technique has been proposed that provides a Minimum Mean Square Error (MMSE) estimate of the desired signal portion in one of the received microphone signals. The procedure is uniquely based on estimates of the second order statistics of the recorded speech signal and the noise signal. A robust speech detection is thus (again) needed. In [12,13], the implementation of the MWF is based on a Generalized Singular Value Decomposition (GSVD) of an input data matrix and a noise data matrix. A cheaper alternative based on a QR Decomposition (QRD) has been proposed in [14]. A subband implementation [15] results in improved intelligibility at a significantly lower cost compared to the fullband approach, making it suitable for e.g. hearing aid applications. In contrast to the GSC, the MWF does not make any a priori assumptions about the signal model so that no or a less severe constraint on the noise sensitivity is needed to guarantee performance when used in combination with small-sized arrays [17,11]. Especially in complicated noise scenarios such as multiple noise sources or diffuse noise, the MWF outperforms the GSC, even with noise sensitivity constraint [11].

The GSC and the MWF both rely on a *speech detection algorithm*. In the GSC case, it is the simultaneous presence of signal model errors and speech detection errors that affects performance. In the MWF case, as the procedure does not require any other a priori information, a reliable speech detection is expected to be crucial to

achieve the potentially better performance. In this paper, we analyze the average effect of speech detection errors on the performance of the GSC (with and without noise sensitivity constraint) and the MWF technique both theoretically and experimentally. Analytical calculations have drawn the attention of many researchers to predict the optimal (i.e. infinite filter length) noise reduction performance achievable by different single and multi-microphone noise reduction systems [18]-[24]. In [18,20,21,23], the theoretically optimal noise reduction performance of the GSC has been derived as a function of the complex coherence functions between the microphone noise signals. In [21,24], a similar theoretical analysis is performed to evaluate a microphone array with postfiltering. In [11], we have developed a theoretical framework to assess the performance of the GSC and the MWF and to evaluate their sensitivity to small signal model errors such as microphone mismatch. In this paper, we extend the theoretical formulas to predict the sensitivity of the GSC (with and without noise sensitivity constraint) and the MWF to speech detection errors. In the simulations and experiments, we focus on the harsh case of small-sized arrays as used in hearing aid applications.

The paper is organized as follows. In Section 2 and Section 3, the concept and the theoretically optimal performance of the GSC and the MWF, respectively, are briefly reviewed. Section 4 theoretically analyzes the average impact of an erroneous speech detection on the performance of the algorithms. It is shown that the MWF technique preserves its benefit over the GSC for a reasonable speech detection error rate of 20% or less, even when the GSC is supplied with a noise sensitivity constraint. In Section 5, the theory is verified based on real recordings made in an office room. The experimental results are found to be well predicted by the theoretical analysis.

Notation

The complex conjugate of a complex value B is denoted as B^*. The superscript T denotes the transpose operator, the superscript H the conjugate transpose.

In the theoretical analysis, we assume stationary signals and infinite filters so that the signals and filters can be represented in the frequency-domain. The microphone signals are $X_k(f)$, $k = 1,\ ...,\ M$ with M the number of microphones. Each microphone signal $X_k(f)$ can be decomposed as

$$X_k(f) = X_k^s(f) + X_k^n(f), \tag{1}$$

where $X_k^s(f)$ denotes the speech component and $X_k^n(f)$ denotes the noise component. Let $\mathbf{X}(f)$ be defined as the stacked data vector

$$\mathbf{X}(f) = \Big[\, X_1(f)\ X_2(f)\ \cdots\ X_M(f) \,\Big]^T. \tag{2}$$

Then, $\mathbf{X}(f)$ can be decomposed as

$$\mathbf{X}(f) = \mathbf{X}^s(f) + \mathbf{X}^n(f), \tag{3}$$

with $\mathbf{X}^s(f)$ and $\mathbf{X}^n(f)$ defined as in (2). During periods of speech + noise, $\mathbf{X}(f) = \mathbf{X}^s(f) + \mathbf{X}^n(f)$. During periods of noise only, $\mathbf{X}(f) = \mathbf{X}^n(f)$.

The Power Spectral Density (PSD) of the signal $X(f)$ is $P_X(f)$, the cross-PSD between signals $X(f)$ and $Y(f)$ is $P_{XY}(f) = \mathcal{E}\{X(f)Y^*(f)\}$. Where needed, the superscripts s and n are used to refer to the contribution of the speech and noise signal only. The noise signal consists of external noise (superscript e) and internal noise (superscript i) e.g. sensor noise, quantization noise, ..., modelled as spatially white noise. We assume a homogenous speech and noise field, i.e., that the PSD of the received speech $P_X^s(f)$, the internal and external noise $P_X^i(f)$ and $P_X^e(f)$ and the received noise $P_X^n(f) = P_X^i(f) + P_X^e(f)$ are the same at each microphone.

2 Generalized Sidelobe Cancellation

2.1 Concept

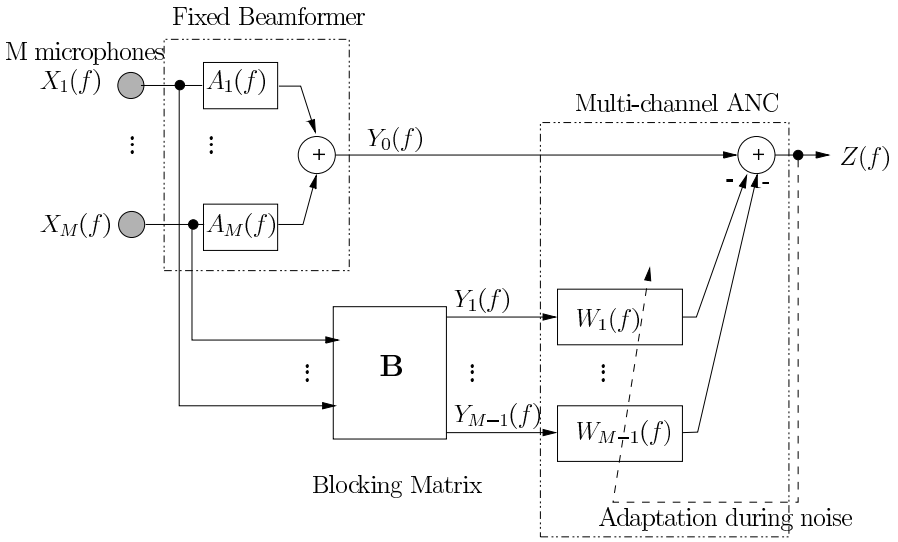


Fig. 1. Concept of the Generalized Sidelobe Canceller.

The Generalized Sidelobe Canceller (GSC) [1–3], depicted in Figure 1, consists of a fixed beamformer, a blocking matrix $\mathbf{B}$ and an Adaptive Noise Canceller (ANC) [25]. The fixed beamformer $\mathbf{A}$ creates the speech reference $Y_0(f) = \mathbf{A}^H\mathbf{X}(f)$ by steering a beam towards the direction of the desired signal. In the sequel, an endfire microphone array is assumed and the desired speaker is assumed to be in the front

at 0 degrees. The $M \times M-1$ blocking matrix $\mathbf{B}$

$$\mathbf{B} = \Big[\mathbf{B}_1(f)\ \mathbf{B}_2(f) \cdots\ \mathbf{B}_{M-1}(f) \Big] \tag{4}$$

creates $M-1$ noise references $Y_j(f) = \mathbf{B}_j^H \mathbf{X}(f)$, $j = 1, \cdots\ M-1$ by steering zeroes to the front. The ANC $\mathbf{W}(f)$

$$\mathbf{W}(f) = \Big[W_1(f)\ W_2(f) \cdots\ W_{M-1}(f) \Big]^H \tag{5}$$

reduces the portion of the noise in the speech reference that is correlated with the noise references. To avoid possible cancellation of the speech, the ANC is adapted during periods of noise only [6,5,7], using a speech/noise detector. We refer to the output of the ANC as $Z(f)$.

In the sequel, a delay-and-sum beamformer is used, i.e.,

$$\mathbf{A}(f) = \frac{1}{M} \Big[1\ \mathrm{e}^{\mathrm{i}2\pi f \frac{d_2 - d_1}{c}} \cdots\ \mathrm{e}^{\mathrm{i}2\pi f \frac{d_M - d_1}{c}} \Big]^H, \tag{6}$$

with $d_j - d_1$ the interspacing between the j-th and the first microphone and c the velocity of sound in air ($c \approx 340$ m/s). For, e.g., $M = 4$, the $M \times M-1$ blocking matrix $\mathbf{B}$ is set to

$$\mathbf{B} = \begin{bmatrix} 1 & -\mathrm{e}^{\mathrm{i}2\pi f \frac{d_2 - d_1}{c}} & 0 & 0 \\ 1 & 0 & -\mathrm{e}^{\mathrm{i}2\pi f \frac{d_3 - d_1}{c}} & 0 \\ 1 & 0 & 0 & -\mathrm{e}^{\mathrm{i}2\pi f \frac{d_4 - d_1}{c}} \end{bmatrix}^H \tag{7}$$

2.2 Theoretical analysis

In [20,23,26], the theoretical (i.e. assuming infinite filter lengths) noise reduction performance NR(f) of the GSC has been expressed in terms of the frequency domain so-called *complex coherence* $\Gamma_{kl}(f)$ between the k-th and l-th microphone signals, $k = 1$, ..., M, $l = 1$, ..., M. In addition to the Power Transfer Function (PTF) of the noise signal $G^n(f) = P_Z^n(f)/P_X^n(f) = 1/\mathrm{NR}(f)$, also the PTF of the desired speech signal $G^s(f) = P_Z^s(f)/P_X^s(f)$ can be expressed as a function of $\Gamma_{kl}(f)$ [11]. In [11], also the PTFs $\tilde{G}^s(f)$ and $\tilde{G}^n(f)$ of the GSC with noise sensitivity constraint have been derived. Based on the PTFs $G^s(f)$ and $G^n(f)$, broadband intelligibility weighted performance measures are computed to assess the overall performance of the algorithm. These results are reviewed in this section.

2.2.1 *Complex coherence function* $\Gamma_{kl}(f)$

The complex coherence $\Gamma_{kl}(f)$ between the k-th and l-th microphone signals with PSD respectively $P_{X_k}(f)$ and $P_{X_l}(f)$ and cross-PSD $P_{X_kX_l}(f)$, is defined as [1]:

$$\Gamma_{kl}(f) = \frac{P_{X_kX_l}}{\sqrt{P_{X_k}P_{X_l}}}. \tag{8}$$

The coherence function Γ_{kl} of a *localized sound source* at an angle θ with respect to the front direction in free and far field equals

$$\Gamma_{kl}(f) = \mathrm{e}^{-\mathrm{i}2\pi f\cos(\theta)\frac{d_k-d_l}{c}} \tag{9}$$

and the coherence Γ_{kl} of a *diffuse sound field* equals

$$\Gamma_{kl}(f) = \frac{\sin(2\pi f(d_k-d_l)/c)}{2\pi f(d_k-d_l)/c} \tag{10}$$

with $d_k - d_l$ the interspacing between microphone k and l [20]. The diffuse sound field is a good approximation of sound fields in strongly reverberant rooms, if the sound sources in the room have no direct path to the microphone array i.e. for a low direct-to-reverberant ratio. The coherence function $\Gamma_{kl}(f)$ due to a set of K *uncorrelated sound sources*, individually inducing coherence functions $\Gamma^j_{kl}(f)$ and PSDs $P^j(f)$, $j = 1, ..., K$ is given by

$$\Gamma_{kl}(f) = \frac{\sum_{j=1}^{K}\alpha_j(f)\Gamma^j_{kl}(f)}{\sum_{j=1}^{K}\alpha_j(f)}, \tag{11}$$

with $\alpha_j(f) = P^j(f)/P^1(f)$. E.g., in the presence of spatially uncorrelated noise, such as sensor noise, the noise coherence $\Gamma^n_{kl}(f)$ is given by [21]

$$\Gamma^n_{kl}(f) = \frac{\Gamma^e_{kl}(f)}{1+\beta(f)} \qquad \text{for } k \neq l, \tag{12}$$

since $\Gamma^i_{kl} = \delta[k-l]$, with Γ^e_{kl} the coherence function of the external noise and with $\beta(f) = P^i_X(f)/P^e_X(f) > 0$ the internal-to-external noise ratio. Typical values for $\beta(f)$ lie between -20 dB and -40 dB [21].

For a homogeneous sound field, the correlation matrix $\mathcal{E}\{\mathbf{X}(f)\mathbf{X}^H(f)\}$ can be expressed as

$$\mathcal{E}\{\mathbf{X}(f)\mathbf{X}^H(f)\} = P_X(f)\mathbf{\Gamma}(f), \tag{13}$$

[1] The frequency parameter f is often omitted in the sequel for the sake of conciseness.

where $\mathbf{\Gamma}(f)$ is the coherence matrix

$$\mathbf{\Gamma}(f) = \begin{bmatrix} 1 & \Gamma_{12}(f) & \cdots & \Gamma_{1M}(f) \\ \Gamma_{21}(f) & 1 & \cdots & \Gamma_{2M}(f) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{M1}(f) & \Gamma_{M2}(f) & \cdots & 1 \end{bmatrix}. \tag{14}$$

Similarly, the speech and the noise correlation matrix equal

$$\mathcal{E}\{\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\} = P_X^s(f)\mathbf{\Gamma}^s(f); \quad \mathcal{E}\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\} = P_X^n(f)\mathbf{\Gamma}^n(f). \tag{15}$$

Remark 1 *Equations (9)-(10) and (13)-(15) assume that the microphones are perfectly matched. Microphone mismatch can be easily incorporated by modifying* $\mathbf{\Gamma}^s(f)$ *and* $\mathbf{\Gamma}^n(f)$ *in (13)-(15) [11]. Suppose that the* k*-th microphone has a gain mismatch* $\Delta\Upsilon(f)$ *and a phase mismatch* $\Delta\Phi(f)$ *(in degrees) at frequency* f *w.r.t. microphone* l*. The matched microphones coherence function* $\Gamma_{kl}(f)$ *should then be replaced by* $\tilde{\Gamma}_{kl}(f)$

$$\tilde{\Gamma}_{kl}(f) = \begin{cases} \Delta\Upsilon(f)\, \mathrm{e}^{j\Delta\Phi(f)\frac{\pi}{180}}\, \Gamma_{kl}(f) & \text{for } l \neq k; \\ |\Delta\Upsilon(f)|^2 \Gamma_{kl}(f) & \text{for } l = k. \end{cases} \tag{16}$$

2.2.2 PTFs $G^s(f)$ and $G^n(f)$ of the GSC

The PTFs $G^s(f)$ and $G^n(f)$ can be expressed as a function of $\mathbf{\Gamma}^s(f)$ and $\mathbf{\Gamma}^n(f)$, the fixed beamformer $\mathbf{A}(f)$ and the blocking matrix $\mathbf{B}(f)$ [11]. Define the stacked noise data vector $\mathbf{Y}^n(f)$ as

$$\mathbf{Y}^n(f) = \begin{bmatrix} Y_1^n(f) & Y_2^n(f) & \cdots & Y_{M-1}^n(f) \end{bmatrix}^T$$

The ANC filters $\mathbf{W}(f)$ are given by [26]:

$$\mathbf{W}(f) = \mathcal{E}\{\mathbf{Y}^n(f)\mathbf{Y}^{n,H}(f)\}^{-1}\mathcal{E}\{\mathbf{Y}^n(f)Y_0^{n,*}(f)\}, \tag{17}$$

where

$$\mathcal{E}\{\mathbf{Y}^n(f)\mathbf{Y}^{n,H}(f)\} = \mathbf{B}^H\mathcal{E}\left\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\right\}\mathbf{B} = P_X^n(f)\mathbf{B}^H\mathbf{\Gamma}^n(f)\mathbf{B} \tag{18}$$

$$\mathcal{E}\{\mathbf{Y}^n(f)Y_0^{n,*}(f)\} = \mathbf{B}^H\mathcal{E}\left\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\right\}\mathbf{A} = P_X^n(f)\mathbf{B}^H\mathbf{\Gamma}^n(f)\mathbf{A}. \tag{19}$$

For a homogeneous noise field, the ANC $\mathbf{W}(f)$ thus only depends on the spatial characteristics of the noise and not on $P_X^n(f)$.

Using (17), the PTFs $G^s(f) = P_Z^s(f)/P_X^s(f)$ and $G^n(f) = P_Z^n(f)/P_X^n(f)$ can be expressed as a function of $\mathbf{\Gamma}^s(f)$ and $\mathbf{\Gamma}^n(f)$ [11]:

$$G^{n|s}(f) = \left(\mathbf{A}^H - \mathbf{W}^H(f)\mathbf{B}^H(f)\right)\mathbf{\Gamma}^{n|s}(f)\left(\mathbf{A} - \mathbf{W}(f)\mathbf{B}(f)\right), \tag{20}$$

where $G^{n|s}(f)$ means $G^n(f)$ or $G^s(f)$, etc. (for compact notation).

2.2.3 PTFs $\tilde{G}^s(f)$ and $\tilde{G}^n(f)$ of the GSC with noise sensitivity constraint

To increase the robustness of the GSC against errors in the assumed signal model, a constraint on the noise sensitivity is often applied [7–9,27]. Especially for small-sized arrays and in complicated noise scenarios, this constraint is found to be crucial to guarantee performance [11]. The *noise sensitivity* is defined as the ratio of the spatially white noise gain to the gain of the desired signal [8,18]. It is used to quantify the sensitivity to errors in the assumed signal model (such as microphone mismatch) as well as to processing errors (such as round off errors due to finite processing accuracy) and determines how much the internal noise, e.g. sensor noise and quantization noise, is amplified. In [8], it is shown that a constraint on the noise sensitivity corresponds to a quadratic inequality constraint on the filters $\mathbf{B}(f)\mathbf{W}(f)$, namely

$$\mathbf{W}^H(f)\mathbf{B}^H(f)\mathbf{B}(f)\mathbf{W}(f) \leq \zeta. \tag{21}$$

In [9], a similar constraint is imposed by inserting uncorrelated noise in the microphone signals[2] used to design the ANC filters $\mathbf{W}(f)$. The unaltered microphone signals are filtered. In [8], it is demonstrated that the method of noise injection is equivalent with a quadratic inequality constraint for a certain level of noise injection and a certain value ζ. Hence, the optimal performance of GSC with noise sensitivity constraint (i.e. the PTFs $\tilde{G}^s(f)$, $\tilde{G}^n(f)$) is obtained by replacing $\mathbf{\Gamma}^n(f)$ in the computation of $\mathbf{W}(f)$ by

$$\mathbf{\Gamma}^n(f) + \eta(f)\mathbf{I}_M, \tag{22}$$

where $\eta(f)$ is the ratio of the injected noise power to the input noise power $P_X^n(f)$ [11] and $\mathbf{I}_M$ is the M-dimensional identity matrix. The equivalent noise injection level $\eta(f)$ for a given ζ is scenario dependent.

2.2.4 Intelligibility weighted performance measures

To assess the effect of the obtained PTFs $G^s(f)$ and $G^n(f)$ on intelligibility, broadband intelligibility weighted measures have been proposed, such as the improve-

[2] Since $\mathbf{B}^H(f)\mathbf{A}(f) = 0$, this corresponds to adding a diagonal loading term to the coherence matrix $\mathbf{\Gamma}^n(f)$ of the noise in the denominator of (17).

ment in *intelligibility weighted Signal-to-Noise Ratio (SNR)* [28], which is defined as

$$\Delta\text{SNR}_{\text{intellig}} = \sum_i I_i(\text{SNR}_{i,\text{out}} - \text{SNR}_{i,\text{in}}), \tag{23}$$

where the band importance function I_i expresses the importance of the i-th one-third octave band with center frequency f_i^{c} for intelligibility, $\text{SNR}_{i,\text{out}}$ is the output SNR (in dB) and $\text{SNR}_{i,\text{in}}$ is the input SNR (in dB) in the i-th one-third octave band. The center frequencies f_i^{c} and the values I_i are defined in [29]; frequencies around 1 kHz are most important while frequencies above 5 kHz are less important for speech intelligibility. Given $G^s(f)$, and $G^n(f)$, the improvement $\Delta\text{SNR}_{\text{intellig}}$ can be calculated using

$$\text{SNR}_{i,\text{in}} = 10\log_{10}\left(\int_{2^{-\frac{1}{6}}f_i^{\text{c}}}^{2^{\frac{1}{6}}f_i^{\text{c}}} P_X^s(f)df \Big/ \int_{2^{-\frac{1}{6}}f_i^{\text{c}}}^{2^{\frac{1}{6}}f_i^{\text{c}}} P_X^n(f)df\right) \tag{24}$$

$$\text{SNR}_{i,\text{out}} = 10\log_{10}\left(\int_{2^{-\frac{1}{6}}f_i^{\text{c}}}^{2^{\frac{1}{6}}f_i^{\text{c}}} P_X^s(f)G^s(f)df \Big/ \int_{2^{-\frac{1}{6}}f_i^{\text{c}}}^{2^{\frac{1}{6}}f_i^{\text{c}}} P_X^n(f)G^n(f)df\right). \tag{25}$$

Similarly, we define an *intelligibility weighted spectral distortion measure* (in dB), called $\text{SD}_{\text{intellig}}$, of the desired speech signal as

$$\text{SD}_{\text{intellig}} = \sum_i I_i \text{SD}_i, \tag{26}$$

with SD_i the average spectral distortion (in dB) in the i-the one-third octave band, calculated as

$$\text{SD}_i = \int_{2^{-\frac{1}{6}}f_i^{\text{c}}}^{2^{\frac{1}{6}}f_i^{\text{c}}} |10\log_{10} G^s(f)|\, df \Big/ \left[\left(2^{\frac{1}{6}} - 2^{-\frac{1}{6}}\right) f_i^{\text{c}}\right]. \tag{27}$$

Summarizing, given the GSC parameters $\mathbf{A}(f)$ and $\mathbf{B}(f)$, the coherence matrices $\mathbf{\Gamma}^s$, $\mathbf{\Gamma}^n$ and the PSD $P_X^s(f)$, $P_X^n(f)$, the intelligibility weighted performance measures can be obtained.

3 Multi-channel Wiener filtering

3.1 Concept

The Multi-channel Wiener Filter (MWF) $\mathbf{w} \in \mathbb{C}^{ML\times 1}$, illustrated in Figure 2, provides a Minimum Mean Square Error (MMSE) estimate $z[k]$ of a delayed version of the (unknown) speech signal $x_i^s[k-\Delta]$ at the i-th (e.g. first) microphone. The

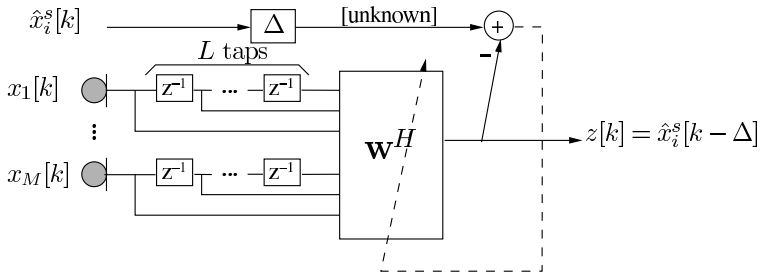


Fig. 2. Concept of the multi-channel Wiener filter.

estimate $z[k]$ is computed as the sum of the M filtered microphone signals, i.e., $z[k] = \hat{x_i^s}[k-\Delta] = \mathbf{w}^H\mathbf{x}[k]$, where

$$\mathbf{w} = \mathcal{E}\{\mathbf{x}[k]\mathbf{x}^H[k]\}^{-1}\mathcal{E}\{\mathbf{x}[k]x_i^{s,*}[k-\Delta]\}, \tag{28}$$

$$\mathbf{x}[k] = [x_1[k]\cdots x_1[k-L+1]\cdots x_M[k]\cdots x_M[k-L+1]]^T \tag{29}$$

where $x_i^s[k-\Delta] = x_i[k-\Delta] - x_i^n[k-\Delta]$ is obviously unknown and where L is the number of filter taps per channel [13].

Assuming that the speech and noise signals are uncorrelated, (28) becomes:

$$\mathbf{w} = \mathcal{E}\{\mathbf{x}[k]\mathbf{x}^H[k]\}^{-1}\left(\mathcal{E}\{\underbrace{\mathbf{x}[k]x_i^*[k-\Delta]}_{\text{speech+noise}}\} - \mathcal{E}\{\underbrace{\mathbf{x}^n[k]x_i^{n,*}[k-\Delta]}_{\text{noise only}}\}\right). \tag{30}$$

The second order statistics of the noise signal are assumed to be sufficiently stationary so that they can be estimated during periods of noise only. A robust speech detection is thus (again) needed. Note that the MWF combines speech distortion and noise reduction in its MMSE optimization criterion

$$\min_{\mathbf{w}} \mathcal{E}\{\left|\mathbf{w}^H\mathbf{x}^s[k] - x_i^s[k-\Delta]\right|^2\} + \mathcal{E}\{\left|\mathbf{w}^H\mathbf{x}^n[k]\right|^2\}. \tag{31}$$

In the sequel, we will assume, without loss of generality, that the speech contribution as received at the first microphone signal is estimated, i.e. $x_1^s[k-\Delta]$.

In [12,13], the MWF is approximated at time instant k by means of a Generalized Singular Value Decomposition (GSVD) of an input and noise data matrix. A cheaper alternative based on a QR Decomposition (QRD) has been proposed in [14]. In [15,16], a subband implementation of the GSVD based algorithm has been developed to increase intelligibility and at the same time reduce complexity. In contrast to the GSC, the MWF does not make any a priori assumptions about the signal model so that it is robust against small signal model errors such as microphone mismatch, reverberation, variations in speaker location and microphone positions [17,12]. It is uniquely based on estimates of the second order statistics of the speech and noise signal.

3.2 Theoretical analysis

In [17,11], the theoretically optimal performance of the MWF, i.e. the PTF of noise and speech as well as the intelligibility weighted performance measures, has been expressed as a function of $\mathbf{\Gamma}(f)$, $\mathbf{\Gamma}^n(f)$ and $\mathbf{\Gamma}^s(f)$. These results are briefly reviewed in this section.

If

$$\mathbf{W}(f) = \Big[W_1(f)\ W_2(f)\ \cdots\ W_M(f) \Big]^H \tag{32}$$

with $W_i(f)$ filtering the i-th microphone signal, $\mathbf{W}(f)$ can be written as [3]:

$$\mathbf{W}(f) = \mathcal{E}\left\{\mathbf{X}(f)\mathbf{X}^H(f)\right\}^{-1} \mathcal{E}\left\{\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\right\}\mathbf{e}_1, \tag{33}$$

where, $\mathbf{e}_1 \in \mathbb{R}^{M\times 1}$ equals the first canonical vector defined as $\mathbf{e}_1 = \Big[1\ 0\ \cdots\ 0 \Big]^T$. The speech correlation matrix $\mathcal{E}\left\{\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\right\}$ is estimated as

$$\mathcal{E}\left\{\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\right\} = \mathcal{E}\left\{\mathbf{X}(f)\mathbf{X}^H(f)\right\} - \mathcal{E}\left\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\right\}, \tag{34}$$

where $\mathcal{E}\left\{\mathbf{X}(f)\mathbf{X}^H(f)\right\}$ is estimated during periods of *speech + noise* and where $\mathcal{E}\left\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\right\}$ is estimated during periods of *noise only*. Using (13)-(15), formula (33) can be rewritten as:

$$\mathbf{W}(f) = \frac{P_X^s(f)}{P_X^s(f) + P_X^n(f)} \mathbf{\Gamma}^{-1}(f)\mathbf{\Gamma}^s(f)\mathbf{e}_1 \tag{35}$$

where, using (11):

$$\mathbf{\Gamma}(f) = \frac{\alpha(f)\mathbf{\Gamma}^n(f) + \mathbf{\Gamma}^s(f)}{1+\alpha(f)}; \tag{36}$$

$$\alpha(f) = \frac{P_X^n(f)}{P_X^s(f)}; \tag{37}$$

In contrast to the GSC, the filter $\mathbf{W}(f)$ not only depends on the spatial characteristics of the noise $\mathbf{\Gamma}^n(f)$, but also on the input SNR $1/\alpha(f)$ and the spatial characteristics of the speech signal $\mathbf{\Gamma}^s(f)$.

Note that the factor $\frac{P_X^s(f)}{P_X^s(f)+P_X^n(f)}$ in (35) corresponds to a *spectral* Wiener filter, which results in additional performance if the spectra of the speech and noise signals differ. Hence, (35) combines spatial and spectral filtering.

[3] In the theoretical analysis, we assume, without loss of generality, that the delay Δ equals zero.

The PTFs $G^n(f) = P_Z^n(f)/P_X^n$ and $G^s(f) = P_Z^s(f)/P_X^s(f)$ from the input $X_1(f)$ to the output $Z(f)$ equal [11]:

$$G^{n|s}(f) = \mathbf{W}^H(f)\mathbf{\Gamma}^{n|s}\mathbf{W}(f). \tag{38}$$

4 Sensitivity to speech detection errors

The theoretical formulas derived in Section 2 and Section 3 allow to predict the average effect of speech detection errors on the performance of the algorithms. In [11], it is shown that the MWF -in contrast to the GSC - does not degrade in the presence of small signal model errors such as microphone mismatch. Especially in complicated noise scenarios, the MWF is found to outperform the GSC as well as the more robust GSC with noise sensitivity constraint [11]. However, like the GSC, the MWF relies on a speech detection algorithm. Since no other, a priori information is used, a reliable speech detection is expected to be crucial to achieve the potentially better performance. In this section, the effect of speech + noise detected as noise only and noise only detected as speech + noise on the performance of the algorithms are both investigated.

The results are illustrated for a small-sized uniform 4-microphone endfire array with microphone interspacing $d = 2$ cm and a diffuse noise field with internal-to-external noise ratio $\beta(f) = -30$ dB. Since in practice, the desired and interfering signals are often speech-like, the improvement $\Delta\text{SNR}_{\text{intellig}}$ is computed using the average long-term speech PSD of the signals found on [30, track 1] for $P_X^s(f)$ and $P_X^n(f)$.

4.1 Generalized Sidelobe Canceller

If the blocking matrix $\mathbf{B}(f)$ of the GSC would be perfect so that the noise references do not contain any speech, the performance of the GSC would be independent of the speech detection. In practice, the a priori assumptions of the GSC are violated due to, e.g., the presence of reverberation, microphone mismatch, ... so that speech leaks into the noise references. To then prevent speech cancellation at high input SNR [4], the ANC is adapted during periods of noise only [5–7] and/or a constraint on the noise sensitivity of the GSC is applied [7,8,27]. For small-sized arrays, the latter is found to be crucial to guarantee performance, even when the ANC is only adapted during noise [11]. Still, in the event of speech leakage, and whenever speech + noise is detected as noise, the ANC may partially cancel the desired speech signal. On the other hand, whenever noise is detected as speech + noise, the performance of the GSC will –apart from a slower convergence rate– not degrade, since the ANC is not adapted during periods of speech.

Suppose that $\delta \times 100\%$ of the samples used by the ANC to compute the noise correlation matrix are actually speech + noise samples. Suppose that the average PSD of the speech in the wrongly detected speech + noise samples equals $\kappa(f)P_X^s(f)$, with $P_X^s(f)$ the average PSD of the speech in the correctly detected *speech + noise* samples. In the simulations, we set $\kappa(f) = 1$ to predict the average impact of speech detection errors. The estimated noise correlation matrix $\mathcal{E}\{\tilde{\mathbf{X}}^n(f)\tilde{\mathbf{X}}^{n,H}(f)\}$ then equals [4]

$$\begin{aligned}\mathcal{E}\{\tilde{\mathbf{X}}^n(f)\tilde{\mathbf{X}}^{n,H}(f)\} &= \mathcal{E}\left\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\right\} + \delta\kappa(f)\mathcal{E}\left\{\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\right\} \\ &= P_X^n(f)\left(\mathbf{\Gamma}^n(f) + \frac{\delta\kappa(f)}{\alpha(f)}\mathbf{\Gamma}^s(f)\right). \end{aligned} \tag{39}$$

The formulas of the ANC filters are then obtained by replacing $\mathcal{E}\left\{\mathbf{Y}^n(f)\mathbf{Y}^{n,H}(f)\right\}$ and $\mathcal{E}\left\{\mathbf{Y}^n(f)Y_0^{n,*}(f)\right\}$ in (17) with

$$\mathcal{E}\left\{\tilde{\mathbf{Y}}^n(f)\tilde{\mathbf{Y}}^{n,H}(f)\right\} = P_X^n(f)\mathbf{B}^H\left(\mathbf{\Gamma}^n + \frac{\delta\kappa(f)}{\alpha(f)}\mathbf{\Gamma}^s\right)\mathbf{B} \tag{40}$$

$$\mathcal{E}\left\{\tilde{\mathbf{Y}}^n(f)\tilde{Y}_0^{n,*}(f)\right\} = P_X^n(f)\mathbf{B}^H\left(\mathbf{\Gamma}^n + \frac{\delta\kappa(f)}{\alpha(f)}\mathbf{\Gamma}^s\right)\mathbf{A}. \tag{41}$$

The impact of an erroneous speech detection on the GSC depends on the quality of the noise references $Y_j(f) = \mathbf{B}_j^H\mathbf{X}(f)$, $j = 1, ..., M-1$: the larger the ratio of speech leakage to noise in the noise references, the larger the impact. From (40)-(41), it follows that the higher the SNR, i.e., the lower $\alpha(f)$, the more the ANC filters and hence the performance will be affected when model errors are present.

4.2 Multi-channel Wiener filtering

4.2.1 Speech + noise samples being erroneously detected as noise only samples

The MWF estimates the correlation matrix $\mathcal{E}\{\mathbf{X}(f)\mathbf{X}^H(f)\}$ during periods of *speech + noise* and the correlation matrix $\mathcal{E}\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\}$ during periods of *noise only*. Assume that the speech and noise signal are uncorrelated and that the second order statistics of the noise vary slowly in time. The speech correlation matrix is $\mathcal{E}\{\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\}$ then estimated using (34). In the case of a perfect speech detection, (34) applies and the MWF is given by (35).

[4] We assume that the second order statistics of the noise are quite stationary, so that the average PSD $P_X^n(f)$ of the noise in the wrongly and correctly detected speech + noise samples equal the average PSD of the noise only samples.

Suppose now that $\delta \times 100\%$ of the samples used to estimate the noise correlation matrix $\mathcal{E}\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\}$ and hence to compute the MWF $\mathbf{W}(f)$ are actually speech + noise samples. Assume that the average PSD of the speech in the $\delta \times 100\%$ wrongly detected speech + noise samples equals $\kappa(f)P_X^s(f)$, with $P_X^s(f)$ the average PSD of the speech in the correctly detected *speech + noise* samples. Then, the estimated noise correlation matrix $\mathcal{E}\{\tilde{\mathbf{X}}^n(f)\tilde{\mathbf{X}}^{n,H}(f)\}$ equals (cf. (39))

$$\mathcal{E}\{\tilde{\mathbf{X}}^n(f)\tilde{\mathbf{X}}^{n,H}(f)\} = P_X^s(f)\left(\alpha(f)\mathbf{\Gamma}^n + \delta\kappa(f)\mathbf{\Gamma}^s\right) \tag{42}$$

instead of $\mathcal{E}\{\mathbf{X}^n(f)\mathbf{X}^{n,H}\}$. Using (34) and (42), the estimated speech correlation matrix $\mathcal{E}\{\tilde{\mathbf{X}}^s(f)\tilde{\mathbf{X}}^{s,H}(f)\}$ becomes

$$\mathcal{E}\{\tilde{\mathbf{X}}^n(f)\tilde{\mathbf{X}}^{n,H}(f)\} = P_X^s(f)(1-\delta\kappa(f))\mathbf{\Gamma}^s(f). \tag{43}$$

Using (42) and (43) in (33), we find that the MWF changes into

$$\mathbf{W}(f) = \frac{(1-\delta\kappa(f))P_X^s(f)}{P_X^s(f)+P_X^n(f)}\mathbf{\Gamma}^{-1}(f)\mathbf{\Gamma}_s(f)\mathbf{e}_1, \tag{44}$$

with $\mathbf{\Gamma}(f)$ and $\mathbf{\Gamma}^s(f)$ the coherence matrices in the case of a perfect speech detection.

Hence, the PTFs $G^s(f)$ and $G^n(f)$ of the MWF are both multiplied by $(1-\delta\kappa(f))^2$. The SNR improvement $G^s(f)/G^n(f)$ at frequency f and hence the improvement $\Delta\text{SNR}_{\text{intellig}}$ remains unchanged, but the speech distortion increases with increasing speech detection error rate δ. The distortion that is caused by the erroneous speech detection, is independent of the input SNR and the noise scenario. Consequently, a higher internal-to-external noise ratio $\beta(f)$ or a constraint on the noise sensitivity will not reduce the additional distortion that is caused by $\delta \neq 0$.

4.2.2 Noise samples being erroneously detected as speech + noise samples

A similar reasoning can be applied when $\delta \times 100\%$ of the samples used to estimate the speech + noise correlation matrix $\mathcal{E}\{\mathbf{X}(f)\mathbf{X}^H(f)\}$ are actually noise only samples. The estimate $\mathcal{E}\{\tilde{\mathbf{X}}(f)\tilde{\mathbf{X}}^H(f)\}$ of the speech + noise correlation matrix and the estimate $\mathcal{E}\{\tilde{\mathbf{X}}^s(f)\tilde{\mathbf{X}}^{s,H}(f)\}$ of the speech correlation matrix then become [5]

$$\begin{aligned}\mathcal{E}\{\tilde{\mathbf{X}}(f)\tilde{\mathbf{X}}^H(f)\} &= \mathcal{E}\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\} + (1-\delta)\mathcal{E}\{\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\} \\ &= P_X^s(f)\left(\alpha(f)\mathbf{\Gamma}^n(f) + (1-\delta)\mathbf{\Gamma}^s(f)\right) \qquad (45)\\ \mathcal{E}\{\tilde{\mathbf{X}}^s(f)\tilde{\mathbf{X}}^{s,H}(f)\} &= (1-\delta)\mathcal{E}\{\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\} = (1-\delta)P_X^s(f)\mathbf{\Gamma}^s(f), \qquad (46)\end{aligned}$$

[5] Since the second order statistics of the noise are assumed to be quite stationary, the average PSD of the noise in the wrongly detected speech + noise samples, the average PSD of the noise in the correctly detected speech + noise samples and the average PSD of the noise only samples are assumed to be the same.

so that the MWF equals

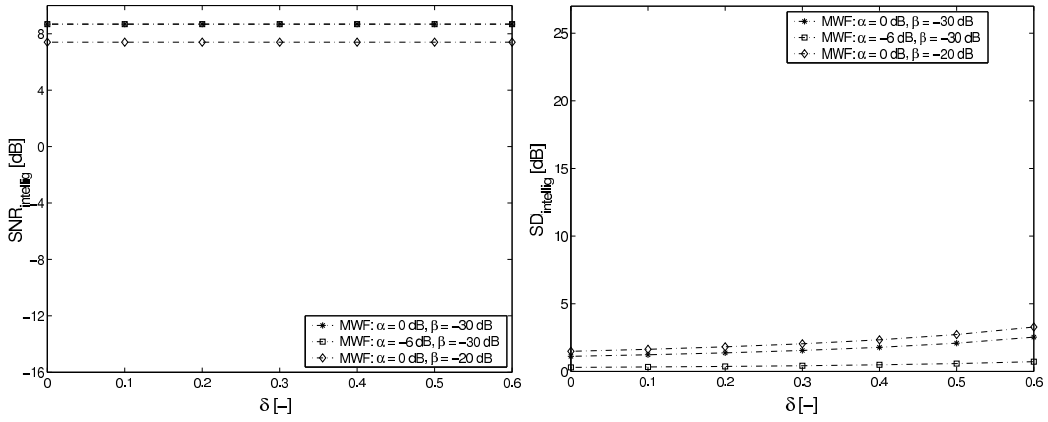
$$\mathbf{W}(f) = \frac{1}{1+\tilde{\alpha}(f)}\tilde{\mathbf{\Gamma}}^{-1}(f)\mathbf{\Gamma}^{s}(f)\mathbf{e}_1, \tag{47}$$

with

$$\tilde{\alpha}(f) = \frac{\alpha(f)}{1-\delta} \tag{48}$$

$$\tilde{\mathbf{\Gamma}}(f) = \frac{\tilde{\alpha}(f)\mathbf{\Gamma}^{n}(f) + \mathbf{\Gamma}^{s}(f)}{1+\tilde{\alpha}(f)}. \tag{49}$$

The MWF acts as if the input PSD of the speech signal and hence the input SNR $1/\alpha(f)$ is decreased by a factor $1-\delta$. As a consequence, the filter will pay more attention to noise reduction than to speech distortion, so that eventually the speech signal is attenuated more than in the case of a perfect speech detection. For an excessive error rate $\delta \times 100\% = 50\%$, the effect corresponds to a decrease in input SNR of only 3 dB, so that the effect of noise detected as speech + noise is expected to be limited.



(a) (b)

Fig. 3. Theoretically optimal performance (ΔSNR$_{\text{intellig}}$ (a) and SD$_{\text{intellig}}$ (b)) of the MWF in a diffuse noise field and in case of a small microphone mismatch, when $(\delta \times 100)$ percent of the samples used to estimate the speech + noise correlation matrix in the MWF are actually noise only samples.

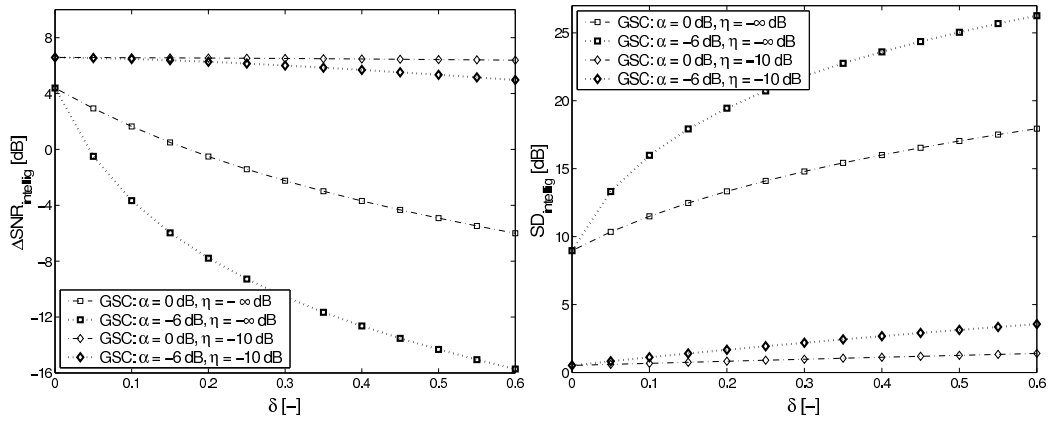
As an illustration, Figure 3(a) and (b) plot the performance of the MWF as a function of δ for two values of internal-to-external noise ratio $\beta(f)$ (i.e. $\beta(f) = -30$ dB and $\beta(f) = -20$ dB) and external noise-to-signal ratio $\alpha(f)$ (i.e. $\alpha(f) = 0$ dB and $\alpha(f) = -6$ dB). In free field, the improvement ΔSNR$_{\text{intellig}}$ is not affected by the erroneous speech detection. The speech distortion SD$_{\text{intellig}}$ slightly grows with increasing δ. The increase in distortion as a function of δ is more significant for scenarios where the output SNR is limited, e.g. due to low input SNR

$1/\alpha(f)$, large$\beta(f)$ and difficult noise scenarios. The effect of noise being detected as speech + noise is less disadvantageous than the effect of speech + noise being detected as noise only. Note that in the latter case, the speech distortion would increase with 1.9 dB for $\delta = 0.2$ and 4.4 dB for $\delta = 0.4$ (see also Figure 5). Hence, if the detection of noise only in a sample is uncertain, the sample should not be used in the computation of the speech + noise correlation matrix in the MWF.

In conclusion, the MWF is especially affected when speech + noise samples are detected as noise-only. In that case, the noise reduction performance of the MWF is preserved, but the speech distortion increases with increasing speech detection error rate.

4.3 Comparison of the GSC and the MWF

In this section, we compare the sensitivity of the GSC and the MWF to speech detection errors. In the simulations, we assume $\kappa(f) = 1$, illustrating the average effect of speech detection errors. The second and the third microphone have a small gain mismatch $\Delta\Upsilon$ of 1 dB and -1 dB w.r.t. the first microphone. In [31], gain and phase differences between microphone characteristics of up to 6 dB and $10°$, respectively, have been reported. In addition, the microphone characteristics may drift over time, making perfect calibration practically impossible.



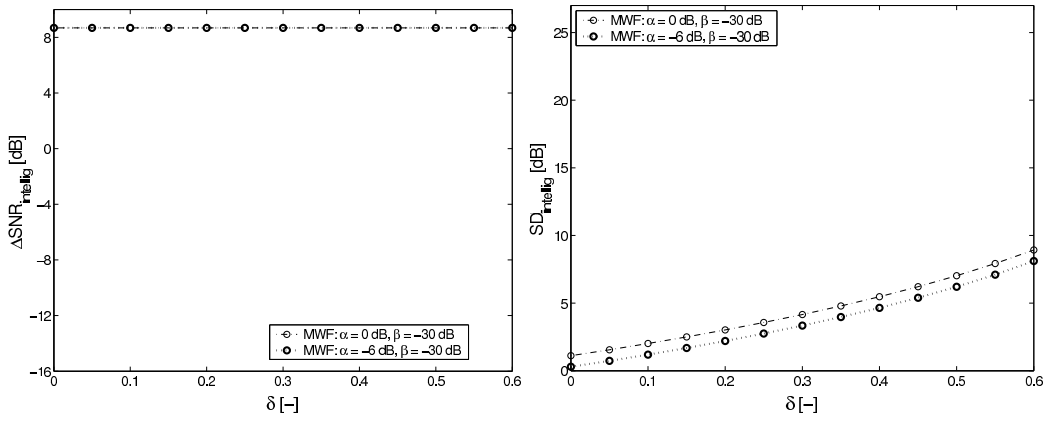
(a) (b)

Fig. 4. Theoretical improvement in $\mathrm{SNR}_{\mathrm{intellig}}$ (a) and spectral distortion $\mathrm{SD}_{\mathrm{intellig}}$ (b) of the GSC in a diffuse noise field with $\beta(f) = -30$ dB and in the case of a small microphone mismatch, when $(\delta \times 100)\%$ of the noise samples used to compute the ANC, actually contain speech.

Figure 4(a) and (b) depict the average impact ($\kappa(f) = 1$) of speech detection errors on the improvement in $\mathrm{SNR}_{\mathrm{intellig}}$ and the speech distortion $\mathrm{SD}_{\mathrm{intellig}}$ of the GSC when $\delta \times 100$ % of the samples used to compute the ANC are actually

speech + noise samples. Two input SNR values $1/\alpha(f)$, i.e. $\alpha(f) = 0\,\text{dB}$ and $\alpha(f) = -6$ dB, are considered. In addition, the performance of the more robust GSC with noise sensitivity constraint (described in Section 2) is depicted for a noise injection level $\eta(f) = -10$ dB. This noise injection level has been found to provide sufficient robustness against gain mismatches $\Delta\Upsilon$ up to ± 3 dB and phase mismatches $\Delta\Phi$ up to $\pm 6^\circ$ for the given microphone array and the given noise scenario [32]. For larger model errors, a more severe constraint will be needed to keep distortion limited, at the expense of less noise reduction.

Without the noise sensitivity constraint, the performance of the GSC is strongly deteriorated by an erroneous speech detection, especially at high input SNR $1/\alpha(f)$. In addition to increasing robustness to model errors [8,11], a noise sensitivity constraint is seen to reduce the impact of speech detection errors. This can be understood as follows. If the model errors are not too large, the noise references of the GSC contain less speech than the speech reference. To cancel the speech effectively the ANC filters have to amplify the speech in the noise references, without excessively amplifying internal noise. Large ANC filter amplitudes, however, result in a high noise sensitivity [8] and thus a large amplification of internal noise. As a consequence, the maximum amplification of the ANC filters and hence the amount of speech cancellation is determined by the internal-to-external noise ratio $\beta(f)$ and the constraint on the noise sensitivity, which is determined by the noise injection level $\eta(f)$.



(a) (b)

Fig. 5. Theoretical improvement in $\text{SNR}_{\text{intellig}}$ (a) and spectral distortion $\text{SD}_{\text{intellig}}$(b) of the MWF in a diffuse noise field with $\beta(f) = -30$ dB and in case of a small microphone mismatch, when $(\delta \times 100)$ percent of the samples used to compute noise correlation matrix in the MWF are actually speech + noise.

Figure 5(a) and Figure 5(b) show the impact of speech detection errors on the improvement $\Delta\text{SNR}_{\text{intellig}}$ and the speech distortion $\text{SD}_{\text{intellig}}$ of the MWF when $\delta \times 100\%$ of the samples used to compute the noise correlation matrix are actually speech + noise samples. Again, two input SNR values $1/\alpha(f)$ are depicted (i.e.,

$\alpha(f) = 0\,\text{dB}$ and $\alpha(f) = -6$ dB). The MWF additionally distorts the speech by $10\log_{10}(1-\delta\kappa(f))^2$ but conserves the improvement $\Delta\text{SNR}_{\text{intellig}}$. The higher the output SNR, the smaller the distortion that is caused by the MWF. In contrast to the GSC, the additional distortion $10\log_{10}(1-\delta\kappa(f))^2$ due to speech detection errors is independent of the input SNR. Compared to the GSC with noise sensitivity constraint, the speech signal is distorted more by the MWF, especially at large error rates δ. For error rates $\delta \times 100\%$ up to 20% the average distortion is limited, while considerably more noise is reduced.

5 Validation through experimental results

The theoretical analysis, presented in Section 4, holds for a localized speech source and localized or diffuse noise sources in the far and free field and assumes infinite filter lengths. In this section, we verify the major conclusions based on real recordings in a reverberant room with reverberation time $T_{60\,\text{dB}} \approx 700$ ms for a speech weighted noise, and when finite filters are used.

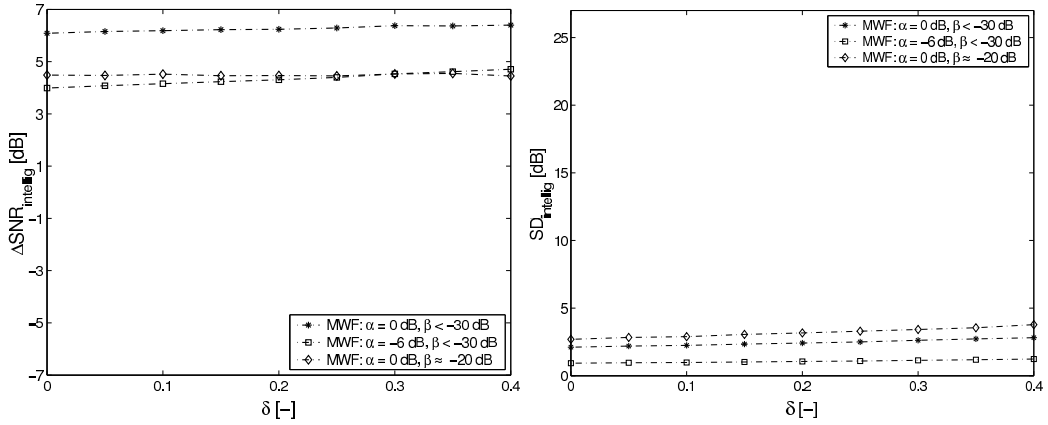
The results are illustrated for a diffuse noise field. A uniform, linear endfire array with 4 omnidirectional microphones (Knowles EM-4368) with interspacing $d = 0.02$ m has been mounted on a dummy head. The desired signal source is positioned at a distance of 1 meter in front of the head. To get an idea of the average performance based on directivity only, stationary speech and noise signals are used with the same, average long-term power spectral density. The signals are colored Gaussian noise with the long-term average spectrum of speech and can be found on [30, track 1]. The speech and noise signal are uncorrelated. The external noise signal has a level of about 70 dB SPL at the center of the head. Since the microphones have an internal noise level of about $25-28$ dB SPL and the external noise signal has the highest energy at low frequencies, the internal-to-external noise level $\beta(f)$ is expected to be smaller than -30 dB at low frequencies. The level of the speech signal is adjusted so that the input SNR at the first microphone equals 0 dB and 6 dB, respectively. During the first 5 seconds only noise is present, during the last 5 seconds the speech and noise signal are both present.

In the experiments, the MWF is implemented in subbands using an oversampled, uniform DFT modulated filterbank with 16 subbands and a downsampling factor of 12 [15,16]. The fixed stage of the GSC has been designed for the 4-microphone array mounted on the dummy head so that –for an anechoic signal coming from $0°$– the energy of the noise references is minimized. The ANC filters are implemented using the Wiener formula to exclude the influence of a potentially slow convergence rate on the overall performance. To improve intelligibility, the microphone signals are pre-emphasized in the GSC prior to processing [33]. At the GSC output, de-emphasis is performed. In the subband GSVD algorithm, the subband implementation intrinsically performs a pre-whitening operation [15]. The filters

of the GSC and the MWF technique are sufficiently long so that a further increase in filter length does not result in a significant increase in performance for the given noise scenario. The ANC of the GSC has 180 taps per noise reference. The subband MWF uses 8 taps per subband and per channel.

The effect of noise being detected as speech + noise and speech + noise being detected as noise only are both investigated. Since the ANC filters of the GSC are adapted during periods of noise only, the former effect is studied only for the MWF. Again, a small gain mismatch of 1 dB and -1 dB has been applied to the second and third microphone w.r.t. the first microphone.

5.1 Noise being erroneously detected as speech + noise



(a) (b)

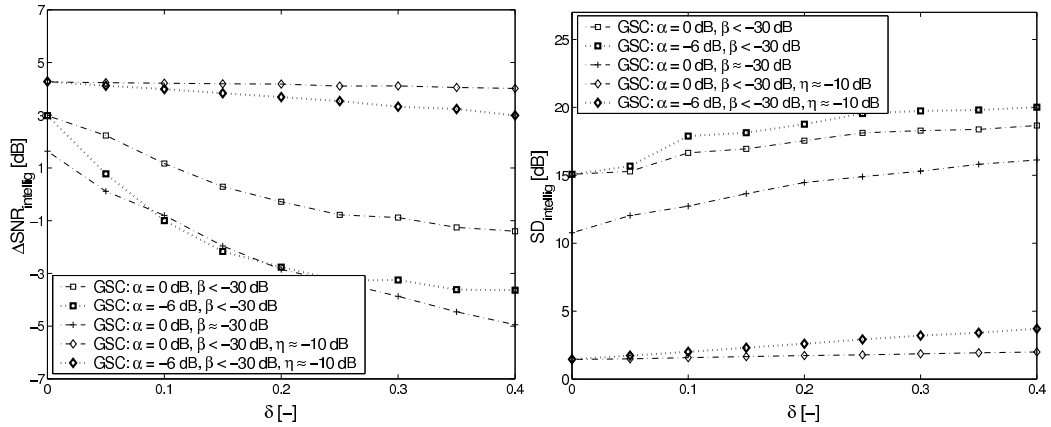
Fig. 6. Effect of noise being erroneously detected as speech + noise on the performance (i.e. $\Delta\text{SNR}_{\text{intellig}}$ (a) and $\text{SD}_{\text{intellig}}$ (b)) of the MWF in a diffuse noise field with internal-to-external noise ratio $\beta(f)$ and input SNR $1/\alpha(f)$; Experimental results.

Figure 6(a) and (b) study the effect of noise only samples detected as speech + noise samples on the performance of the MWF. The performance measures $\Delta\text{SNR}_{\text{intellig}}$ and $\text{SD}_{\text{intellig}}$ are depicted as a function of δ with $\delta \times 100\%$ the percentage of samples that are detected as speech + noise but actually contain noise only. Two input noise-to-signal ratios $\alpha(f)$ are considered: $\alpha(f) = 0$ dB and $\alpha(f) = -6$ dB. The influence of a higher internal-to-external noise ratio $\beta(f) \approx -20$ dB is shown too. The behavior of the curves is quite well predicted by Figure 3: an increased error rate δ only leads to a small increase in distortion $\text{SD}_{\text{intellig}}$. The increase in distortion $\text{SD}_{\text{intellig}}$ grows with increasing $\beta(f)$ and increasing $\alpha(f)$. The noise reduction performance $\Delta\text{SNR}_{\text{intellig}}$ is not significantly affected by the speech detection errors. In contrast to the results of Figure 3, the improvement $\Delta\text{SNR}_{\text{intellig}}$ drops for decreasing α. This difference is due to the presence of reverberation in the speech signal. The MWF trades off reduction of diffuse noise and cancellation of the reverberant

part in the speech signal. The filter will pay more attention to noise reduction for large $\alpha(f)$. The improvement $\Delta\text{SNR}_{\text{intellig}}$ also slightly increases as a function of δ since a larger δ corresponds to an increase of $\alpha(f)$ (see (47)).

5.2 Speech + noise being erroneously detected as noise only

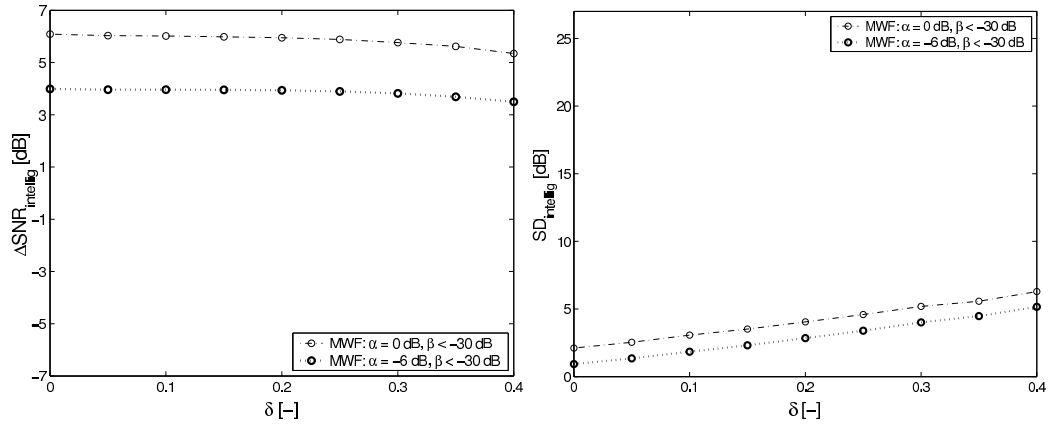
5.2.1 GSC



(a) (b)

Fig. 7. Performance of the GSC without and with noise sensitivity constraint (noise injection level $\eta(f) \approx -10$ dB) in a diffuse noise field with input SNR $= 1/\alpha(f)$ when $(\delta \times 100)\%$ of the samples used to compute the noise correlation matrix are actually speech + noise samples; Experimental results. (a) Improvement $\Delta\text{SNR}_{\text{intellig}}$; (b) Spectral distortion $\text{SD}_{\text{intellig}}$.

Figure 7(a) and Figure 7(b) plot the improvement $\Delta\text{SNR}_{\text{intellig}}$ and the spectral distortion $\text{SD}_{\text{intellig}}$, respectively, of the GSC when $\delta \times 100\%$ of the samples used to compute the filters, actually contain speech + noise. Two input signal-to-noise ratios $1/\alpha(f)$ are depicted: $\alpha = 0$ dB and $\alpha = -6$ dB. As has been predicted by Figure 4 in Section 4, the improvement $\Delta\text{SNR}_{\text{intellig}}$ of the GSC is strongly affected by an erroneous speech detection, especially for high input SNR. Compared to Figure 4, the distortion $\text{SD}_{\text{intellig}}$ at $\delta = 0$ is worse due to a lower internal-to-external noise ratio $\beta(f) < -30$ dB at low frequencies and the presence of reverberation. Indeed, the experimentally found performance of the GSC with $\beta(f)$ increased to about -30 dB is more similar to the theoretical predicted one in Figure 4. In Figure 7, the performance of the more robust GSC with a noise sensitivity constraint is also depicted for a noise injection level $\eta(f) \approx -10$ dB. The drastic impact of speech detection errors on the GSC is then tremendously reduced.



(a) (b)

Fig. 8. Performance of the MWF in a diffuse noise field with input SNR $= 1/\alpha(f)$ when $(\delta \times 100)\%$ of the samples used to compute the noise correlation matrix are actually speech + noise samples; Experimental results. (a) Improvement $\Delta\text{SNR}_{\text{intellig}}$; (b) Spectral distortion $\text{SD}_{\text{intellig}}$.

5.2.2 Multi-channel Wiener filter

Figure 8(a) and (b) illustrate the effect of an erroneous speech detection on the performance of the MWF. The improvement in $\text{SNR}_{\text{intellig}}$ of the MWF is hardly affected by speech detection errors, while the average speech distortion $\text{SD}_{\text{intellig}}$ that is due to erroneous speech detection increases with about $|10\log_{10}(1-\delta)^2|$, independently of the input SNR. In contrast to Figure 5, the improvement $\Delta\text{SNR}_{\text{intellig}}$ depends again on α due to reverberation: a lower α results in less noise reduction and less speech distortion.

Already in the presence of small model errors, the MWF is less sensitive to an erroneous speech detection than the GSC without noise sensitivity constraint. Compared to the more robust GSC with noise sensitivity constraint, the MWF reduces more noise, especially at low input SNR, while the distortion is limited for $\delta < 0.2$. In addition, in the presence of larger model errors the performance of the GSC with noise sensitivity constraint will further degrade. The MWF on the other hand, is insensitive to model errors [11].

6 Conclusions

We have analyzed the sensitivity to speech detection errors of the GSC and a MWF technique both theoretically and experimentally. Since the latter does not rely on any a priori information about the signal model, a reliable speech detection is expected to be crucial to achieve its potentially better performance. In Section 2 and

Section 3 formulas for the optimal performance of both algorithms have been derived.

In Section 4, the formulas have been extended to assess the average effect of speech detection errors on the performance. In the simulations, we focussed on small-sized arrays as, e.g., used in hearing aid applications. Theoretical analysis showed that the noise reduction performance of the MWF is –in contrast to the GSC– not degraded by speech detection errors, but the speech gets distorted, when speech + noise samples are detected as noise only samples. The distortion is proportional to $1/(1-\delta)$, where $\delta \times 100\%$ equals the percentage of samples that are detected as noise only but actually contain speech + noise. This average distortion is independent of the noise scenario and the input SNR and for error rates up to 20%, the distortion is limited. Already in the presence of small model errors, the MWF is found to be less sensitive to an erroneous speech detection than the GSC, despite their only dependence on a good performing speech detection. A constraint on the noise sensitivity of the GSC is not only essential to increase robustness to small signal model errors when small-sized arrays are used, but also to limit the drastic impact of speech detection errors. Depending on the noise sensitivity constraint and the size of the model error, the robust GSC with noise sensitivity constraint distorts the speech signal less than the MWF, especially at large δ. However, compared to the MWF, less noise is reduced. In short, the MWF preserves its potential benefit over the GSC for a reasonable speech detection error rate of 20% or less, even when the GSC is supplied with a noise sensitivity constraint.

The theoretical analysis assumes infinite filter lengths and a localized speech source and localized or diffuse noise sources in the far and free field. In Section 5, the theoretical results have been validated for real recordings in an office room and finite filter lengths. The experimental results confirm the theory.

Acknowledgements

Ann Spriet is a Research Assistant supported by the F.W.O. Vlaanderen. This research work was carried out at the ESAT laboratory and Lab. Exp. ORL of the K. U. Leuven, in the frame of the Belgian State, Prime Minister's Office - Federal Office for Scientific, Technical and Cultural Affairs - Interuniversity Attraction Pole (2002-2007) - IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling'), the Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology) of the Flemish Government, Research Project FWO nr. G.0233.01 ('Signal processing and automatic patient fitting for advanced auditory prostheses') and was partially sponsored by Cochlear. The scientific responsibility is assumed by its authors.

References

[1] L. J. Griffiths, C. W. Jim, An alternative approach to linearly constrained adaptive beamforming, IEEE Trans. Antennas Propag. 30 (1) (1982) 27–34.

[2] B. D. Van Veen, K. M. Buckley, Beamforming: A versatile approach to spatial filtering, IEEE ASSP Magazine 5 (2) (1988) 4–24.

[3] K. M. Buckley, Broad-band beamforming and the Generalized Sidelobe Canceller, IEEE Trans. Acoust., Speech, and Signal Processing 34 (5) (1986) 1322–1323.

[4] P. M. Peterson, A.-M. Wie, W. M. Rabinowitz, P. M. Zurek, Robustness of an adaptive beamforming method for hearing aids, Acta Otolaryngol. Suppl. 469 (1990) 85–90.

[5] J. E. Greenberg, P. M. Zurek, Evaluation of an adaptive beamforming method for hearing aids, Journal of the Acoustical Society of America 91 (3) (1992) 1662–1676.

[6] D. Van Compernolle, Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings, in: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2, Albuquerque, 1990, pp. 833–836.

[7] O. Hoshuyama, A. Sugiyama, A. Hirano, A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters, IEEE Trans. Signal Processing 47 (1999) 2677–2683.

[8] H. Cox, R. M. Zeskind, M. M. Owen, Robust Adaptive Beamforming, IEEE Trans. Acoust., Speech, and Signal Processing 35 (10) (1987) 1365–1376.

[9] N. K. Jablon, Adaptive beamforming with the Generalized Sidelobe Canceller in the presence of array imperfections, IEEE Trans. Antennas Propag. 34 (1986) 996–1012.

[10] M. W. Hoffman, T. D. Trine, K. M. Buckley, D. J. Van Tasell, Robust adaptive microphone array processing for hearing aids: realistic speech enhancement predictions, Journal of the Acoustical Society of America 96 (1994) 759–770.

[11] A. Spriet, M. Moonen, J. Wouters, Robustness analysis of multi-channel Wiener filtering and Generalized Sidelobe Cancellation for multi-microphone noise reduction in hearing aid applications, *Accepted for publication in* IEEE Trans. Speech, Audio Processing .

[12] S. Doclo, M. Moonen, GSVD-Based Optimal Filtering for Multi-Microphone Speech Enhancement, Springer-Verlag, 2001, Ch. 6 in "Microphone Arrays: Signal Processing Techniques and Applications" (Brandstein, M. S. and Ward, D. B., Eds.), pp. 111–132.

[13] S. Doclo, M. Moonen, GSVD-Based Optimal Filtering for Single and Multimicrophone Speech Enhancement, IEEE Trans. Signal Processing 50 (9) (2002) 2230–2244.

[14] G. Rombouts, M. Moonen, QRD-based optimal filtering for acoustic noise reduction, in: Proc. European Signal Processing Conf. (EUSIPCO), Vol. 3, Toulouse, France, 2002, pp. 301–304.

[15] A. Spriet, M. Moonen, J. Wouters, A multi-channel subband generalized singular value decomposition approach to speech enhancement, European Transactions on Telecommunications 13 (2) (2002) 149–158.

[16] A. Spriet, M. Moonen, J. Wouters, A multichannel subband GSVD based approach for speech enhancement in hearing aids, in: Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC), Darmstadt, Germany, 2001, pp. 187–191.

[17] A. Spriet, M. Moonen, J. Wouters, Robustness analysis of GSVD based optimal filtering and Generalized Sidelobe Canceller for hearing aid applications, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 2001, pp. 31–34.

[18] P. M. Peterson, Adaptive array processing for multiple microphone hearing aids, Ph.D. thesis, Dept. Elect. Eng. and Comp. Sci., M.I.T, Cambridge, MA, available as Res. Lab. Elect. Techn. Rept. 541 (1989).

[19] J. Bitzer, K. U. Simmer, K.-D. Kammeyer, Multichannel noise reduction - Algorithms and theoretical limits-, in: Proc. European Signal Processing Conf. (EUSIPCO), Vol. 1, Rhodes, Greece, 1999, pp. 105–108.

[20] J. Bitzer, K. U. Simmer, K.-D. Kammeyer, Theoretical noise reduction limits of the Generalized Sidelobe Canceller (GSC) for speech enhancement, in: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Vol. 5, Phoenix, Arizona, USA, 1999, pp. 2965–2968.

[21] J. Bitzer, K. U. Simmer, K.-D. Kammeyer, Multi-microphone noise reduction techniques as front-end devices for speech recognition, Speech Communication 34 (2001) 3–12.

[22] S. Gannot, B. D., W. E., Theoretical analysis of the general transfer function GSC, in: Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC), Darmstadt, Germany, 2001, pp. 103–106.

[23] S. E. Nordholm, Y. H. Leung, Performance limits of the broadband Generalized Sidelobe Cancelling structure in an isotropic noise field, Journal of the Acoustical Society of America 107 (2) (2000) 1057–1060.

[24] C. Marro, Y. Mahieux, K. U. Simmer, Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering, IEEE Trans. Speech, Audio Processing 6 (3) (1998) 240–259.

[25] B. Widrow, S. Stearns, Adaptive Signal Processing, Prentice Hall, Englewood Cliffs, 1985.

[26] J. Bitzer, K. U. Simmer, K.-D. Kammeyer, An alternative implementation of the superdirective beamformer, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz NY, USA, 1999, pp. 7–10.

[27] M. W. Hoffman, K. M. Buckley, Robust time-domain processing of broadband acoustic data, IEEE Trans. Speech, Audio Processing 3 (1995) 193–203.

[28] J. E. Greenberg, P. M. Peterson, P. M. Zurek, Intelligibility-weighted measures of speech-to-interference ratio and speech system performance, Journal of the Acoustical Society of America 94 (5) (1993) 3009–3010.

[29] Acoustical Society of America, ANSI S3.5-1997 American National Standard Methods for calculation of the speech intelligibility index (Jun. 1997).

[30] International Collegium of Rehabilitative Audiology, Noise Signals ICRA, Compact Disc (Ver 0.3 1997).

[31] L. B. Jensen, Hearing aid with adaptive matching of input transducers, U.S. patent, 6741714 (May 2004).

[32] A. Spriet, M. Moonen, J. Wouters, Robustness analysis of multi-channel Wiener filtering and Generalized Sidelobe Cancellation for multi-microphone noise reduction in hearing aid applications, Tech. Rep. ESAT-SISTA/TR 02-81a, ESAT-SCD/SISTA, KU Leuven (Belgium), available at ftp://ftp.esat.kuleuven.ac.be/pub/sista/spriet/reports/02-81a.pdf (Sep. 2002).

[33] M. J. Link, K. M. Buckley, Prewhitening for intelligibility gain in hearing aid arrays, Journal of the Acoustical Society of America 93 (4) (1993) 2139–2140.

List of Figures

1 Concept of the Generalized Sidelobe Canceller. 4

2 Concept of the multi-channel Wiener filter. 10

3 Theoretically optimal performance ($\Delta\text{SNR}_{\text{intellig}}$ (a) and $\text{SD}_{\text{intellig}}$ (b)) of the MWF in a diffuse noise field and in case of a small microphone mismatch, when $(\delta \times 100)$ percent of the samples used to estimate the speech + noise correlation matrix in the MWF are actually noise only samples. 15

4 Theoretical improvement in $\text{SNR}_{\text{intellig}}$ (a) and spectral distortion $\text{SD}_{\text{intellig}}$ (b) of the GSC in a diffuse noise field with $\beta(f) = -30$ dB and in the case of a small microphone mismatch, when $(\delta \times 100)\%$ of the noise samples used to compute the ANC, actually contain speech. 16

5 Theoretical improvement in $\text{SNR}_{\text{intellig}}$ (a) and spectral distortion $\text{SD}_{\text{intellig}}$(b) of the MWF in a diffuse noise field with $\beta(f) = -30$ dB and in case of a small microphone mismatch, when $(\delta \times 100)$ percent of the samples used to compute noise correlation matrix in the MWF are actually speech + noise. 17

6 Effect of noise being erroneously detected as speech + noise on the performance (i.e. $\Delta\text{SNR}_{\text{intellig}}$ (a) and $\text{SD}_{\text{intellig}}$ (b)) of the MWF in a diffuse noise field with internal-to-external noise ratio $\beta(f)$ and input SNR $1/\alpha(f)$; Experimental results. 19

7 Performance of the GSC without and with noise sensitivity constraint (noise injection level $\eta(f) \approx -10$ dB) in a diffuse noise field with input SNR $= 1/\alpha(f)$ when $(\delta \times 100)\%$ of the samples used to compute the noise correlation matrix are actually speech + noise samples; Experimental results. (a) Improvement $\Delta\text{SNR}_{\text{intellig}}$; (b) Spectral distortion $\text{SD}_{\text{intellig}}$. 20

8 Performance of the MWF in a diffuse noise field with input SNR $= 1/\alpha(f)$ when $(\delta \times 100)\%$ of the samples used to compute the noise correlation matrix are actually speech + noise samples; Experimental results. (a) Improvement $\Delta\text{SNR}_{\text{intellig}}$; (b) Spectral distortion $\text{SD}_{\text{intellig}}$. 21