Katholieke Universiteit Leuven

Departement Elektrotechniek ESAT-SISTA/TR 10-242

A VAD-robust Multichannel Wiener Filter algorithm for noise reduction in hearing aids[1]

Bram Cornelis[2] , Marc Moonen[2], Jan Wouters[3]

Proc. of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, May 2011, pp. 281-284

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.ac.be* in the directory *pub/sista/bcorneli/reports/ICASSP11.pdf*. (c) 2011 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), Kasteelpark Arenberg 10, 3001 Leuven, Belgium. Tel. +32 16 321797, Fax +32 16 321970, WWW: *http://www.esat.kuleuven.ac.be/sista*, E-mail: *bram.cornelis@esat.kuleuven.ac.be*. Bram Cornelis is funded by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). This research work was carried out at the ESAT Laboratory of Katholieke Universiteit Leuven in the frame of the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007-2011), Concerted Research Action GOA-MaNet and research project FWO nr. G.0600.08 ('Signal processing and network design for wireless acoustic sensor networks'). The scientific responsibility is assumed by its authors.

[3]K.U.Leuven, Dept. of Neurosciences, ExpORL, Herestraat 49/721, 3000 Leuven, Belgium.

Abstract

The Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) is a promising multi-microphone noise reduction technique, in particular for hearing aid applications. Its benefit over other techniques has been shown in several theoretical and experimental contributions. In theoretical studies, a single target speech source is commonly assumed, as this facilitates the analysis. In this contribution, we first prove that an algorithm, that implicitly assumes a single target speech source, is also more robust against estimation errors in the speech second order statistics, compared to a standard SDW-MWF algorithm. Secondly, as any SDW-MWF algorithm relies on a voice activity detector (VAD), a novel VAD-robust extension is also proposed. It is shown theoretically and through experiments with a realistic VAD that the new algorithm indeed achieves a good performance, even at low input SNR's where the VAD error rate is high.

A VAD-ROBUST MULTICHANNEL WIENER FILTER ALGORITHM FOR NOISE REDUCTION IN HEARING AIDS

Bram Cornelis[1,*], *Marc Moonen*[1], *Jan Wouters*[2]

[1] Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10, 3001 Leuven, Belgium,
[2] Katholieke Universiteit Leuven, ExpORL, Herestraat 49/721, 3000 Leuven, Belgium.

ABSTRACT

The Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) is a promising multi-microphone noise reduction technique, in particular for hearing aid applications. Its benefit over other techniques has been shown in several theoretical and experimental contributions. In theoretical studies, a single target speech source is commonly assumed, as this facilitates the analysis. In this contribution, we first prove that an algorithm, that implicitly assumes a single target speech source, is also more robust against estimation errors in the speech second order statistics, compared to a standard SDW-MWF algorithm. Secondly, as any SDW-MWF algorithm relies on a voice activity detector (VAD), a novel VAD-robust extension is also proposed. It is shown theoretically and through experiments with a realistic VAD that the new algorithm indeed achieves a good performance, even at low input SNR's where the VAD error rate is high.

Index Terms— binaural hearing aids, noise reduction, multichannel Wiener filtering, voice activity detection (VAD)

1. INTRODUCTION

Noise reduction has been an active area of research for many years, with applications in hearing aids, hands-free communications and teleconferencing. The Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) [1] is a promising multi-microphone technique for speech in noise scenarios, in particular for hearing aid applications. The SDW-MWF does not require prior knowledge about the target speech signal location and microphone characteristics, unlike fixed beamformers or adaptive beamformers such as the generalized sidelobe canceller (GSC) [2]. As a result, the SDW-MWF is more robust against imperfections such as microphone mismatch [3]. Like the GSC, the SDW-MWF relies on a voice activity detection (VAD) algorithm which classifies frames as either speech+noise or noise-only frames. For a moderate VAD error rate, the SDW-MWF achieves a better performance compared to a GSC-type procedure [4]. If a wireless link is available for exchanging signals between a left and a right hearing aid, the SDW-MWF is again an excellent choice for such so-called binaural hearing aids, as the localization cues can be preserved in addition to achieving a better noise reduction performance [5].

*Bram Cornelis is funded by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). This research work was carried out at the ESAT Laboratory of Katholieke Universiteit Leuven in the frame of the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007-2011), Concerted Research Action GOA-MaNet and research project FWO nr. G.0600.08 ('Signal processing and network design for wireless acoustic sensor networks'). The scientific responsibility is assumed by its authors.

In theoretical studies on the (frequency-domain) SDW-MWF (e.g. [3–6]), it is commonly assumed that there is a single target speech source, so that the speech correlation matrix is of rank one. As a consequence, the SDW-MWF can be decomposed into a spatial filter followed by a single-channel Wiener postfilter. In addition to facilitating a theoretical analysis, this structure is also conceptually interesting, as for example, the spatial filter and postfilter could be updated at different rates, or extended independently with other features.

In this paper, we will illustrate that SDW-MWF algorithms, which implicitly assume a single target source and that decompose the overall filter as a spatial filter followed by a single-channel postfilter, have some additional benefits over the standard SDW-MWF algorithm. First, it is proven that the rank one SDW-MWF algorithms are more robust to estimation errors in the speech second order statistics. Secondly, it is shown that these rank-one algorithms are inherently insensitive to VAD errors if the noise is stationary. Finally, for realistic nonstationary noise environments, an extension is proposed which improves the SNR performance at low input SNR's, where the VAD error rate is high.

The paper is organized as follows: in section 2, the notation and model is introduced and the different SDW-MWF algorithms are reviewed. In section 3, a theoretical analysis of the impact of estimation errors in the speech second order statistics is given. In section 4, an extension to the rank-one SDW-MWF algorithm is proposed, which further increases the robustness to estimation errors and VAD errors at low input SNR's. Simulations of the new algorithm in a realistic acoustic environment, and using a real VAD, are presented in section 5. Finally, conclusions are given in section 6.

2. NOTATION AND MULTICHANNEL WIENER FILTER

2.1. Notation and correlation matrix estimation

We consider a microphone array consisting of N microphones. The nth microphone signal $Y_n(\omega)$ can be specified in the frequency domain as

$$Y_n(\omega) = X_n(\omega) + V_n(\omega), \quad n = 1 \dots N, \tag{1}$$

where $X_n(\omega)$ represents the speech component and $V_n(\omega)$ represents the noise component in the nth microphone. For conciseness, we will omit the frequency variable ω from now on. The signals Y_n, X_n and V_n are stacked in the N-dimensional vectors $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{v}$, with $\mathbf{y}{=}\mathbf{x}+\mathbf{v}$. The correlation matrix $\mathbf{R}_y$, the speech correlation matrix $\mathbf{R}_x$ and the noise correlation matrix $\mathbf{R}_v$ are then defined as

$$\mathbf{R}_y = \mathcal{E}\{\mathbf{y}\mathbf{y}^H\}, \;\; \mathbf{R}_x = \mathcal{E}\{\mathbf{x}\mathbf{x}^H\}, \;\; \mathbf{R}_v = \mathcal{E}\{\mathbf{v}\mathbf{v}^H\}\,, \tag{2}$$

where $\mathcal{E}$ denotes the expected value operator. In order to make a distinction between speech+noise and noise-only frames, a voice activity detection (VAD) algorithm is used. The correlation matrix estimates $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_v$ are then recursively updated (per frequency bin) as:

$$\widehat{\mathrm{SNR}}_{\mathrm{out}} \quad = \quad \rho - \frac{\rho\rho^{''} - P_s|\rho^{'}|^2}{\frac{\mu^2}{(\mu+\rho)^2}\left[P_s|A_{\mathrm{ref}}|^2\rho + P_s A_{\mathrm{ref}}(\rho^{'}) + P_s A^*_{\mathrm{ref}}(\rho^{'})^* + \rho^{''} + \frac{2\mu+\rho}{\mu^2}(\rho\rho^{''} - P_s|\rho^{'}|^2)\right]} \tag{11}$$

- In speech+noise frames:

$$\hat{\mathbf{R}}_y[m+1] = \lambda_y \hat{\mathbf{R}}_y[m] + (1-\lambda_y)\mathbf{y}[m+1]\mathbf{y}^H[m+1] \ , \tag{3}$$
$$\hat{\mathbf{R}}_v[m+1] = \hat{\mathbf{R}}_v[m] \ ,$$

- In noise-only frames:

$$\hat{\mathbf{R}}_y[m+1] = \hat{\mathbf{R}}_y[m] \ ,$$
$$\hat{\mathbf{R}}_v[m+1] = \lambda_v \hat{\mathbf{R}}_v[m] + (1-\lambda_v)\mathbf{v}[m+1]\mathbf{v}^H[m+1] \ . \tag{4}$$

λ_y and λ_v are forgetting factors (usually chosen close to 1), and m is the frame-index. For conciseness, the frame-index will be omitted from now on. Assuming that the speech and the noise components are uncorrelated and that the noise is moderately stationary, the speech correlation matrix can be found as $\hat{\mathbf{R}}_x$=$\hat{\mathbf{R}}_y - \hat{\mathbf{R}}_v$.

The goal of the noise reduction procedure is to minimize the error between the output signal Z=$\mathbf{w}^H\mathbf{y}$ and the (unknown) speech component $X_{\mathrm{ref}} = \mathbf{e}_{\mathrm{ref}}^H\mathbf{x}$ in the reference microphone signal. Vector $\mathbf{e}_{\mathrm{ref}} = [0 \ldots 0 \ \ 1 \ \ 0 \ldots 0]^T$ is an N-dimensional vector where the entry corresponding to the reference microphone is equal to one.

2.2. Multichannel Wiener Filter (MWF) algorithms

The Multichannel Wiener Filter (MWF) produces a minimum-mean-square-error (MMSE) estimate of the speech component in the reference microphone. To provide a more explicit tradeoff between speech distortion and noise reduction, the Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) has been proposed, which minimizes a weighted sum of the residual noise energy and the speech distortion energy [1]:

$$\mathbf{w}_{\mathrm{SDW-MWF}} = \left(\mathbf{R}_x + \mu\mathbf{R}_v\right)^{-1}\mathbf{R}_x\mathbf{e}_{\mathrm{ref}} \ . \tag{5}$$

The trade-off parameter μ allows putting more emphasis on noise reduction, at the cost of a higher speech distortion.

In the case of a single target speech source, the speech signal vector can be modeled as $\mathbf{x} = \mathbf{a}S$, where the N-dimensional steering vector $\mathbf{a}$ contains the acoustic transfer functions from the speech source to the microphones (including room acoustics, microphone characteristics and head shadow effect) and S denotes the speech signal. The speech correlation matrix is then a rank-one matrix, i.e.

$$\mathbf{R}_x = P_s\mathbf{a}\mathbf{a}^H \ , \tag{6}$$

with $P_s = \mathcal{E}\{|S|^2\}$ the power of the speech signal. The SDW-MWF (5) then reduces to the following filter:

$$\mathbf{w} = \mathbf{R}_v^{-1}\mathbf{a} \ . \ \frac{P_s \ A^*_{\mathrm{ref}}}{\mu + \rho} \ , \tag{7}$$

with $A^*_{\mathrm{ref}} = \mathbf{a}^H\mathbf{e}_{\mathrm{ref}}$ and $\rho = P_s\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}$. The narrowband output SNR (i.e. per frequency bin) obtained with this filter is also equal to ρ. Expression (7) shows that the filter has the conceptually interesting structure of a spatial filter ($\mathbf{R}_v^{-1}\mathbf{a}$) followed by a single-channel Wiener postfilter ($\frac{P_s \ A^*_{\mathrm{ref}}}{\mu+\rho}$). Unfortunately expression (7) does not allow for an easy practical algorithm as the speech power and steering vector would have to be somehow estimated or else calibrated beforehand.

In [6], it is shown that an alternative (but theoretically equivalent) SDW-MWF formula can be derived, again assuming a rank-one speech correlation matrix (denoted here as rank-one MWF or R1-MWF), which only depends on the speech and noise second order statistics, i.e.

$$\mathbf{w}_{R1-MWF} = \mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\mathrm{ref}} \ . \ \frac{1}{\mu + \mathrm{Tr}\{\mathbf{R}_v^{-1}\mathbf{R}_x\}} \ , \tag{8}$$

where $\mathrm{Tr}\{.\}$ is the trace operator. Another alternative was originally proposed as distortionless multichannel Wiener filter in [7] and is denoted as spatial prediction MWF (SP-MWF) in [8]:

$$\mathbf{w}_{\mathrm{SP-MWF}} = \mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\mathrm{ref}} \ . \ \frac{\mathbf{e}_{\mathrm{ref}}^H\mathbf{R}_x\mathbf{e}_{\mathrm{ref}}}{\mu\mathbf{e}_{\mathrm{ref}}^H\mathbf{R}_x\mathbf{e}_{\mathrm{ref}} + \mathbf{e}_{\mathrm{ref}}^H\mathbf{R}_x\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\mathrm{ref}}} \ . \tag{9}$$

In contrast to (7), expressions (8) and (9) only rely on the speech and noise second order statistics, which are relatively easy to estimate. They are however still structured as a spatial filter followed by a single-channel postfilter, where only the postfilter depends on μ, in contrast to (5).

3. IMPACT OF SPEECH SECOND ORDER STATISTICS ESTIMATION ERRORS

The impact of estimation errors in the speech second order statistics is now investigated for a scenario with a single target speech source. Inaccurate estimation of the speech statistics occurs because of several reasons [3]. The speech and noise may be nonstationary, while $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_v$ are estimated at different moments in time. Speech detection errors made by the VAD will also introduce estimation errors in both the speech and the noise correlation matrices.

In the case of a single speech source, we assume the estimated speech correlation matrix $\hat{\mathbf{R}}_x$ is equal to (up to a scalar factor):

$$\hat{\mathbf{R}}_x = P_s\mathbf{a}\mathbf{a}^H + \mathbf{\Delta} \ , \tag{10}$$

i.e. $\hat{\mathbf{R}}_x$ will be equal to the theoretical rank-one matrix plus a full rank (Hermitian) error matrix $\mathbf{\Delta}$. By plugging (10) into the filter expressions (5), (8) and (9), the narrowband output SNR (i.e. per frequency bin) can be calculated.

It can be shown that the rank-one algorithms based on (8) and (9) always obtain a higher narrowband output SNR than the standard SDW-MWF (5). In particular, the narrowband output SNR of the standard SDW-MWF can be shown to be equal to expression (11) shown at the top of this page [8]. In this formula, $\rho^{'} = \mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{\Delta}\mathbf{e}_{\mathrm{ref}}$ and $\rho^{''} = \mathbf{e}_{\mathrm{ref}}^H\mathbf{\Delta}^H\mathbf{R}_v^{-1}\mathbf{\Delta}\mathbf{e}_{\mathrm{ref}}$. The obtained output SNR with estimated $\hat{\mathbf{R}}_x$ is thus equal to the optimal output SNR ($= \rho$) minus a positive bias term. This bias term is moreover dependent on μ, and it can be shown that the narrowband output SNR monotonically increases as μ increases [8]. While in theory, for a single target speech source, μ only occurs in the single-channel postfilter (formula (7)) and thus has no effect on the narrowband output SNR, in practice a high μ value has to be chosen when using the standard SDW-MWF (5). At higher μ values, more speech distortion is however introduced.

For the R1-MWF and SP-MWF, it can be shown that the obtained output SNR is independent of μ and equal to [8]:

$$\widehat{\mathrm{SNR}}_{\mathrm{out}} = \rho - \frac{\rho\rho^{''} - P_s|\rho^{'}|^2}{P_s|A_{\mathrm{ref}}|^2\rho + P_s A_{\mathrm{ref}}(\rho^{'}) + P_s A^*_{\mathrm{ref}}(\rho^{'})^* + \rho^{''}} \ . \tag{12}$$

With the standard SDW-MWF, this output SNR can only be obtained for the limit case ($\mu \rightarrow \infty$) as can be calculated from (11). For realistic values of μ, the standard SDW-MWF is therefore always outperformed by the R1-MWF and SP-MWF.

4. VAD-ROBUST R1-MWF

4.1. VAD robustness under stationary noise

We now focus on the spatial filter part ($\mathbf{R}_v^{-1}\mathbf{a}$) of (7). If the estimated filter is parallel to this filter, the optimal narrowband SNR performance can be achieved. We hereby assume that the single-channel postfilter can compensate for any arbitrary scaling factor so that the speech distortion is limited. First, for any values α and β with $\alpha \neq \beta$, and for a single target speech source, we find that (where $\cong$ means "is parallel to"):

$$\begin{aligned}\mathbf{R}_v^{-1}\mathbf{a} &\cong \mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\text{ref}} \cong (\mathbf{R}_v + \alpha\mathbf{R}_x)^{-1}(\beta-\alpha)\mathbf{R}_x\mathbf{e}_{\text{ref}} \ , \\ &\cong (\mathbf{R}_v + \alpha\mathbf{R}_x)^{-1}\big[(\mathbf{R}_v + \alpha\mathbf{R}_x) + (\beta-\alpha)\mathbf{R}_x\big]\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}} \ , \\ &\cong (\mathbf{R}_v + \alpha\mathbf{R}_x)^{-1}(\mathbf{R}_v + \beta\mathbf{R}_x)\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}} \ . \end{aligned} \tag{13}$$

This implies that if we have any two different mixes of $\mathbf{R}_v$ and $\mathbf{R}_x$ as in (13), a filter which is parallel to the optimal spatial filter can be obtained.

When the noise is stationary, VAD errors in the estimated noise and speech+noise correlation matrices can be modelled as [4]:

$$\hat{\mathbf{R}}_v = \mathbf{R}_v + \delta_1\mathbf{R}_x \quad \text{and} \quad \hat{\mathbf{R}}_y = \mathbf{R}_v + \delta_2\mathbf{R}_x \ , \tag{14}$$

so that the spatial filter part of the R1-MWF and SP-MWF is estimated as (using $\hat{\mathbf{R}}_x = \hat{\mathbf{R}}_y - \hat{\mathbf{R}}_v$):

$$\begin{aligned}\hat{\mathbf{w}} &= \hat{\mathbf{R}}_v^{-1}\hat{\mathbf{R}}_y\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}} \ , &(15)\\ &= (\mathbf{R}_v + \delta_1\mathbf{R}_x)^{-1}(\mathbf{R}_v + \delta_2\mathbf{R}_x)\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}} \ , &(16)\end{aligned}$$

which, by (13), is parallel to $\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{e}_{\text{ref}}$. For stationary noise, we thus see that the spatial filter part of (8) and (9), remarkably, is robust against VAD errors.

4.2. VAD robustness under nonstationary noise

For (spectrally) nonstationary noise, (14) is inaccurate, as the changing noise power will lead to different weighting factors on $\mathbf{R}_v$, i.e.

$$\hat{\mathbf{R}}_v = \gamma_1\mathbf{R}_v + \delta_1\mathbf{R}_x \quad \text{and} \quad \hat{\mathbf{R}}_y = \gamma_2\mathbf{R}_v + \delta_2\mathbf{R}_x \ . \tag{17}$$

As a result, (13) is no longer valid and the R1-MWF will not be parallel to the optimal filter. If $\hat{\mathbf{R}}_y$ can be scaled so that $\gamma_1 \approx \gamma_2$, the optimal performance is however again achieved. We therefore search a scalar c so that $\hat{\mathbf{R}}_v - c\hat{\mathbf{R}}_y \cong \mathbf{R}_x$. As $\mathbf{R}_x$ is a rank-one matrix, the rank of $\hat{\mathbf{R}}_v - c\hat{\mathbf{R}}_y$ has to be minimized, which can be done by trace minimization [9]. This results in the following convex optimization problem:

$$\min_c \ \text{Tr}\big\{|\hat{\mathbf{R}}_v - c\hat{\mathbf{R}}_y|\big\} \ , \tag{18}$$

where $|.|$ is the absolute value operator. Instead of (18), we will however minimize the following convex problem,

$$\min_c \ \text{Tr}\big\{(\hat{\mathbf{R}}_v - c\hat{\mathbf{R}}_y)^2\big\} \ , \tag{19}$$

as the optimum can then be found analytically as

$$c_{\text{opt}} = \frac{\text{vec}(\hat{\mathbf{R}}_v)^H\text{vec}(\hat{\mathbf{R}}_y)}{\text{vec}(\hat{\mathbf{R}}_y)^H\text{vec}(\hat{\mathbf{R}}_y)} \ , \tag{20}$$

where $\text{vec}(.)$ stacks the columns of a matrix into a column vector.

The scaling factor obtained in (20) can be wrong if the input SNR is too high, because the trace minimization will then lead to a c for which $\hat{\mathbf{R}}_v - c\hat{\mathbf{R}}_y \cong \mathbf{R}_v$. In that case the normal R1-MWF solution (without scaling of $\hat{\mathbf{R}}_y$) is better. Fortunately, at higher input SNR's, there are also less VAD errors so that a good performance is achieved. As a switching criterion, we propose (with $\|.\|$ the matrix 2-norm):

if $1 - \text{TOL} \leq \frac{\|c_{\text{opt}}\hat{\mathbf{R}}_y\|}{\|\hat{\mathbf{R}}_v\|} \leq 1 + \text{TOL}$ **then**
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{R}}_v^{-1}(c_{\text{opt}}\hat{\mathbf{R}}_y)\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}}$
else
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{R}}_v^{-1}\hat{\mathbf{R}}_y\mathbf{e}_{\text{ref}} - \mathbf{e}_{\text{ref}}$
end if

5. SIMULATIONS

5.1. Setup

We consider a binaural setup of two behind-the-ear hearing aids connected by a wireless link. There are two omnidirectional microphones per device, and we assume that the link allows for transmitting one audiosignal (i.e. the front microphone signal) to the other device (full-duplex). The noise reduction procedure therefore has access to a total of $N = 3$ microphone signals per device.

Head-related transfer functions (HRTF's) are measured in a reverberant room (SII-weighted reverberation time of 0.62s) on a Cortex MK2 manikin. We consider three acoustical scenario's: S0N45, S315N90 and S0N90/180/270, where the target speech (S) and interfering noise (N) source(s) are positioned at the specified azimuthal angles (with 0° in front of the head, 90° to the right of the head).

As speech stimulus, 4 sentences of the Dutch VU sentence material [10] were used. Multitalker babble noise was used as interfering noise signal(s). To assess the performance, the speech intelligibility weighted SNR improvement [11] is calculated on the second part of the signals (last two sentences), to allow the filters to converge.

The standard SDW-MWF (5), R1-MWF (8) and VAD-robust R1-MWF (section 4.2), were implemented in a weighted overlap-add (WOLA) filterbank. The signals are sampled at $\text{f}_s = 20480Hz$, segmented in frames of 128 samples with 75% overlap, and windowed by a Hann Window. Other used parameters are: $\mu = 5$, $\lambda_y = \lambda_v = 0.9995$, $\text{TOL} = 0.05$.

As VAD algorithm, we make a decision fusion (using the wireless link) of log-energy-based VAD algorithms calculated at both sides of the head, together with a cross-correlation based VAD (which assumes a signal from 0° is target speech). The decision fusion rule is based on local SNR estimates as in [12]. We note that two of the scenarios pose a difficulty for the cross-correlation VAD: in S315N90 the speech location deviates from 0°, in S0N90/180/270 one of the noise sources (at 180°) also appears to be from 0°.

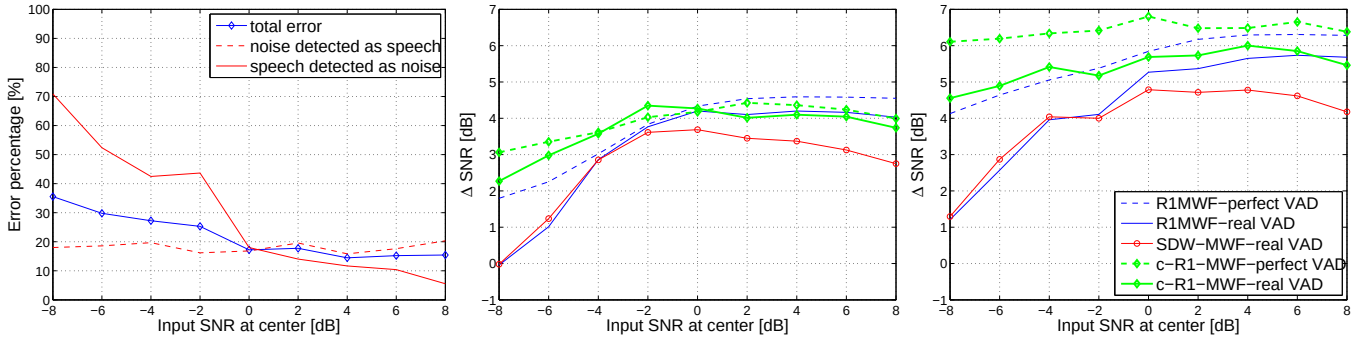
5.2. Results

In the figure, the VAD performance and left and right SNR improvements (versus the front microphone signals) are shown for input SNR's ranging from -8dB to +8dB (measured in absence of the head). The results show that the R1-MWF indeed outperforms the standard SDW-MWF as was explained in section 3, especially at higher input SNR's (where $\mathbf{R}_x$ has more weight in the SDW-MWF formula and imperfections have more impact). The VAD-robust MWF of section 4.2, denoted as c-R1-MWF, generally leads to a better performance than the R1-MWF, even if a perfect VAD is used.[1] The improvement of the c-R1-MWF is retained when using a real VAD, especially at lower SNR's (high VAD error rate). On the other hand no performance is lost at higher input SNR's as the algorithm then switches back to the normal R1-MWF solution.

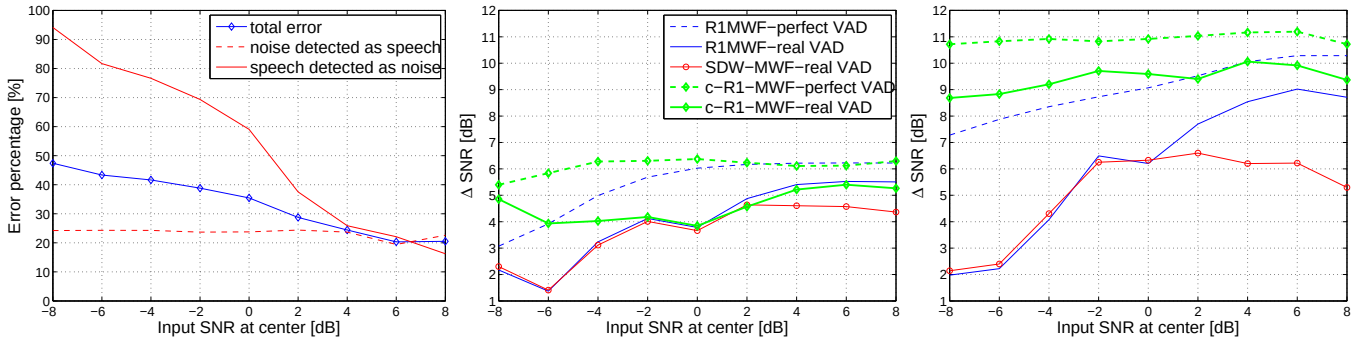
6. CONCLUSION

In this paper, we have proven theoretically and shown through experiments on a binaural hearing aid setup that SDW-MWF algorithms,

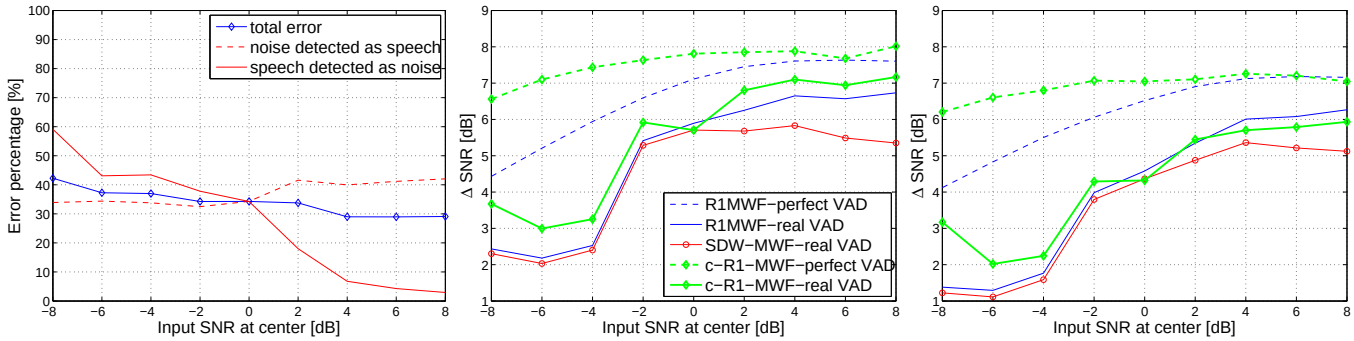
[1] For the perfect VAD, a binary decision is made for every frame (for the entire frequency range). As a result, even for a perfect VAD, $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{R}}_v$ can be badly scaled at the frequencies with a low average input SNR (where speech is mostly absent), which can however be improved by the c-R1-MWF.



(a) **S0 N45**: VAD performance (b) **S0 N45**: Left output (c) **S0 N45**: Right output



(d) **S315 N90**: VAD performance (e) **S315 N90**: Left output (f) **S315 N90**: Right output



(g) **S0 N90/180/270**: VAD performance (h) **S0 N90/180/270**: Left output (i) **S0 N90/180/270**: Right output

which implicitly assume a single target speech source and are structured as a spatial filter followed by a single-channel postfilter, are more robust to estimation errors in the speech second order statistics and hence outperform the standard SDW-MWF algorithm. In addition, a novel extension was proposed which further increases the robustness against estimation errors and VAD errors at low input SNR's.

7. REFERENCES

[1] Doclo, S. and Moonen, M., "GSVD-based optimal filtering for single and multi-microphone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sept. 2002.

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

[3] A. Spriet, M. Moonen, and J. Wouters, "Robustness Analysis of Multi-channel Wiener Filtering and Generalized Sidelobe Cancellation for Multi-microphone Noise Reduction in Hearing Aid Applications," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 487–503, July 2005.

[4] A. Spriet, M. Moonen, and J. Wouters, "The impact of speech detection errors on the noise reduction performance of multi-channel Wiener filtering and Generalized Sidelobe Cancellation," *Signal Process.*, vol. 85, no. 6, pp. 1073–1088, June 2005.

[5] B. Cornelis, S. Doclo, T. Van den Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[6] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.

[7] J. Benesty, J. Chen, and Y. Huang, *Noncausal (frequency-domain) optimal filters*, chapter 6 in "Microphone Array Signal Processing", pp. 115–137, Springer-Verlag, 2008.

[8] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1368–1381, July 2011.

[9] M. Fazel, H. Hindi, and S. Boyd, "Rank minimization and applications in system theory," in *Proc. Amer. Control Conf.*, June 2004, vol. 4, pp. 3273 – 3278.

[10] N. J. Versfeld, L. Daalder, J. M. Festen, and T. Houtgast, "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," *J. Acoust. Soc. Amer.*, vol. 107, no. 3, pp. 1671–1684, 2000.

[11] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *J. Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.

[12] V. Berisha, H. Kwon, and S. Spanias, "Real-time implementation of a distributed voice activity detector," in *IEEE Workshop on Sensor Array and Multichannel Process.*, July 2006, pp. 659–662.