VARIABLE SPEECH DISTORTION WEIGHTED MULTICHANNEL WIENER FILTER BASED ON SOFT OUTPUT VOICE ACTIVITY DETECTION FOR NOISE REDUCTION IN HEARING AIDS

Kim Ngo[1]*, Ann Spriet*[1,2]*, Marc Moonen*[1]*, Jan Wouters*[2] *and Søren Holdt Jensen*[3]

[1]Katholieke Universiteit Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[2]Katholieke Universiteit Leuven, ExpORL, O. & N2, Herestraat 49/721, B-3000 Leuven, Belgium
[3]Aalborg University, Dept. Electronic Systems, Niels Jernes Vej 12, DK-9220 Aalborg, Denmark

ABSTRACT

This paper presents a variable Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) based on soft output Voice Activity Detection (VAD) which is used for noise reduction in hearing aids. A traditional SDW-MWF uses a fixed parameter to trade-off between noise reduction and speech distortion. Consequently, the improvement in noise reduction comes at the cost of a higher speech distortion. With a variable SDW-MWF the goal is to improve the noise reduction without increasing the speech distortion. A soft output VAD is used to distinguish between speech, noise and to incorporate a variable trade-off. In speech dominant segments it is desirable to have less noise reduction to avoid speech distortion. In noise dominant segments it is desirable to have as much noise reduction as possible. Experimental results with a variable SDW-MWF show a SNR improvement with a lower speech distortion compared to a SDW-MWF.

Index Terms— Multichannel Wiener Filter, Noise reduction, Speech distortion, Soft output VAD.

1. INTRODUCTION

Background noise (multiple speakers, traffic etc.) is a significant problem for hearing aid users and is especially damaging to speech intelligibility. To overcome this problem both single-channel and multichannel noise reduction schemes have been proposed. The limitation of single-channel noise reduction is that only temporal and spectral signal characteristics are used. Multichannel noise reduction in addition exploits the spatial diversity of the speech and the noise signals. The objective of noise reduction algorithms is to maximally reduce the noise while minimizing speech distortion. A known multichannel noise reduction technique is the Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF) [1] [2] that allows for a trade-off between noise reduction and speech distortion. However, the improvement in noise reduction comes at the cost of a higher speech distortion. Recently, soft output Voice Activity Detection (VAD) has been used in speech enhancement for gain modification and noise spectrum estimation [3] [4] [5]. The concept is to increase the gain when there is a high probability that speech is present and to apply a lower gain in the presence of noise i.e. when there is a lower probability that speech is present. Soft output VAD has also been used for controlling the compression gain when noise reduction and dynamic range compression are integrated [6]. Here, the soft output VAD was used to distinguish between the speech and the noise dominant segments in order not to amplify the residual noise after the noise reduction. This paper presents a variable SDW-MWF based on soft output VAD which allows for a variable trade-off between noise reduction and speech distortion in the SDW-MWF procedure.

The paper is organised as follows. In Section 2 the signal model and the SDW-MWF is described. In Section 3 the SDW-MWF is extended with a soft output VAD. In Section 4 experimental results are presented. The work is summarized in Section 5.

2. MULTICHANNEL WIENER FILTER

2.1. Signal model

Let $X_i(f), i = 1, ..., M$ denote the frequency-domain microphone signals

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

where f is the frequency domain variable and the superscripts s and n are used to refer to the speech and the noise contribution of a signal, respectively. Let $\mathbf{X}(f) \in \mathbb{C}^{M\times 1}$ be defined as the stacked vector

$$\mathbf{X}(f) = [X_1(f)\ X_2(f)\ ...\ X_M(f)]^T \tag{2}$$
$$= \mathbf{X}^s(f) + \mathbf{X}^n(f) \tag{3}$$

This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven, in the frame of the Marie-Curie Fellowship EST-SIGNAL program (http://est-signal.i3s.unice.fr) under contract No. MEST-CT-2005-021175, and the Concerted Research Action GOA-AMBioRICS. Ann Spriet is a postdoctoral researcher funded by F.W.O.-Vlaanderen. The scientific responsibility is assumed by its authors.

where the superscript T denotes the transpose. Defining $H_i^s(f)$ as the acoustic transfer function from the speech source $S(f)$ to the *i*-th microphone, $\mathbf{X}^s(f)$ can be written as

$$\mathbf{X}(f) = \mathbf{H}(f)S(f) + \mathbf{X}^n(f) \quad (4)$$

$$\mathbf{X}^s(f) = \mathbf{H}^s(f)S(f) = \tilde{\mathbf{H}}^s(f)X_1^s(f) \quad (5)$$

with $\tilde{\mathbf{H}}^s(f)$ the vector with transfer function ratios relative to the first microphone.

In addition, we define the noise and the speech correlation matrix as

$$\mathbf{R}^n(f) = \varepsilon\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\} \quad (6)$$

$$\mathbf{R}^s(f) = \varepsilon\{\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\} = P_{X_i}^s(f)\tilde{\mathbf{H}}^s(f)\tilde{\mathbf{H}}^{s,H}(f) \quad (7)$$

where $\varepsilon\{\}$ denotes the expectation operator, H denotes Hermitian transpose and $P_{X_i}^s(f)$ is the power spectral density (PSD) of the speech in the *i*-th microphone signal.

The MWF optimally estimates a desired signal, based on a Minimum Mean Square Error (MMSE) criterion. Here, the desired signal is the speech component $X_1^s(f)$ in the first microphone. The MWF has been extended to the SDW-MWF that allows for a trade-off between noise reduction and speech distortion using a trade-off parameter μ [1] [2]. Assume that the speech and the noise signals are statistically independent, then the optimal SDW-MWF that provides an estimate of the speech component in the first microphone is given by

$$\mathbf{W}(f) = \left(\mathbf{R}^s(f) + \mu\mathbf{R}^n(f)\right)^{-1}\mathbf{R}^s(f)\mathbf{e}_1 \quad (8)$$

where the $M \times 1$ vector $\mathbf{e}_1$ equals the first canonical vector defined as $\mathbf{e}_1 = [1 \quad 0 \quad ... \quad 0]^T$. The second-order statistics of the noise are assumed to be stationary which means $\mathbf{R}^s(f)$ can be estimated as $\mathbf{R}^s(f) = \mathbf{R}^x(f) - \mathbf{R}^n(f)$ where $\mathbf{R}^x(f)$ and $\mathbf{R}^n(f)$ are estimated during periods of speech+noise and periods of noise-only, respectively. For μ = 1 the SDW-MWF solution reduces to the MWF solution which for $\mu > 1$ the residual noise level will be reduced at the cost of a higher speech distortion. The output $Z(f)$ of the SDW-MWF can then be written as

$$Z(f) = \mathbf{W}^H(f)\mathbf{X}(f). \quad (9)$$

3. MULTICHANNEL WIENER FILTER WITH SOFT OUTPUT VAD

Traditionally, the trade-off parameter μ is set to a fixed value and the improvement in noise reduction comes at the cost of a higher speech distortion. Furthermore, the speech+noise segments and the noise-only segments are weighted equally, whereas it is desirable to have more noise reduction in the noise-only segments compared to the speech+noise segments. With a variable SDW-MWF it is possible to distinguish between the speech+noise segments and noise-only segments using a soft output VAD. The soft output VAD can be implemented according to [3] [4] [5]. The variable SDW-MWF is derived from the MSE criterion as (The frequency parameter f is omitted in the sequel for the sake of conciseness)

$$\mathbf{W} = \arg\min_{\mathbf{W}} \varepsilon\{|X_1^s - \mathbf{W}^H\mathbf{X}|^2\} \quad (10)$$

$$\mathbf{W} = \arg\min_{\mathbf{W}} \varepsilon\{p \cdot |X_1^s - \mathbf{W}^H\mathbf{X}|^2 + (1-p) \cdot |\mathbf{W}^H\mathbf{X}^n|^2\} \quad (11)$$

where p is the probability that speech is present in a given signal segment. The solution is then given by

$$\mathbf{W} = (p \cdot \varepsilon\{\mathbf{X}^s\mathbf{X}^{s,H}\} + p \cdot \varepsilon\{\mathbf{X}^n\mathbf{X}^{n,H}\} + (1-p) \cdot \varepsilon\{\mathbf{X}^n\mathbf{X}^{n,H}\})^{-1}p \cdot \varepsilon\{\mathbf{X}^s X_1^{s,H}\} \quad (12)$$

$$\mathbf{W} = (p \cdot \varepsilon\{\mathbf{X}^s\mathbf{X}^{s,H}\} + \varepsilon\{\mathbf{X}^n\mathbf{X}^{n,H}\})^{-1}p \cdot \varepsilon\{\mathbf{X}^s X_1^{s,H}\} \quad (13)$$

the variable SDW-MWF can then be written as

$$\mathbf{W} = \left(\mathbf{R}^s + \left(\tfrac{1}{p}\right)\mathbf{R}^n\right)^{-1}\mathbf{R}^s\mathbf{e}_1. \quad (14)$$

Compared to Eq. 8 with the fixed μ the term $\frac{1}{p}$ is now changing based on the soft output VAD. The concept goes as follows

- If $p = 0$, i.e. the probability that speech is presence is zero, the variable SDW-MWF will attenuate the noise by applying $\mathbf{W} \leftarrow 0$.
- If $p = 1$ the variable SDW-MWF solution corresponds to the MWF solution.
- If $0 < p < 1$ there is a trade-off between noise reduction and speech distortion.

3.1. Spatial and Spectral Filtering

For further analysis the SDW-MWF can be decomposed into a spatial filter and a spectral filter [7] [8]. Assuming that $\mathbf{R}^s$ is rank 1 and using the definitions in Eq. 7 we can write the optimal filter as

$$\mathbf{W} = \left(P_{X_1}^s\tilde{\mathbf{H}}^s\tilde{\mathbf{H}}^{s,H} + \left(\tfrac{1}{p}\right)\mathbf{R}^n\right)^{-1}P_{X_1}^s\tilde{\mathbf{H}}^s. \quad (15)$$

Applying the matrix inversion lemma the optimal filter can then be decomposed into

$$\mathbf{W} = \underbrace{\frac{\mathbf{R}^{n^{-1}}\tilde{\mathbf{H}}^{s}}{\tilde{\mathbf{H}}^{s,H}\mathbf{R}^{n^{-1}}\tilde{\mathbf{H}}^{s}}}_{\text{TF-GSC}} \underbrace{\left(\frac{P^{s}_{X_1}}{P^{s}_{X_1} + P^{n}_{X_1}}\right)}_{\text{Postfilter}} \tag{16}$$

where

$$P^{n}_{X_1} = \frac{1}{\tilde{\mathbf{H}}^{s,H}(\frac{1}{p})\mathbf{R}^{n^{-1}}\tilde{\mathbf{H}}^{s}} \tag{17}$$

is the output noise power from the Transfer Function Generalized Sidelobe Canceller (TF-GSC) beamformer. This shows that the residual noise after the beamformer (spatial filter) can be further suppressed by the postfilter (spectral filter). The beamformer reduces the noise while keeping the speech component in the first microphone signal undistorted. The soft output VAD $\frac{1}{p}$ only affects the spectral post filtering. The postfilter can be viewed as a single-channel Wiener filter where each frequency component is attenuated based on the signal-to-noise ratio.

4. EXPERIMENTAL RESULTS

In this Section, experimental results for the variable SDW-MWF based on soft output VAD are presented and compared to SDW-MWF with fixed values for μ.

4.1. Set-up and performance measures

We have performed simulations with a 2-microphone behind-the-ear hearing aid. The speech is located at 0° and the two multi-talker babble noise sources are located at 120° and 180°.

To assess the noise reduction performance the intelligibility-weighted signal-to-noise ratio (SNR) [9] is used which is defined as

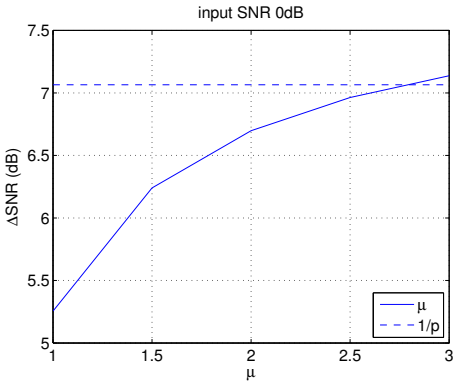
$$\Delta SNR_{intellig} = \sum_i I_i (SNR_{i,out} - SNR_{i,in}) \tag{18}$$

where I_i is the band importance function defined in [10] and $\text{SNR}_{i,out}$ and $\text{SNR}_{i,in}$ represents the output SNR and the input SNR (in dB) of the *i*-th band, respectively. For the speech distortion an intelligibility weighted spectral distortion measure is used defined as

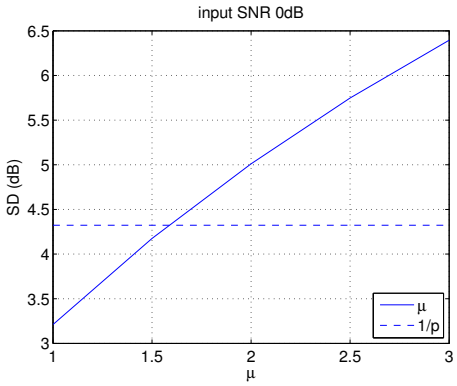
$$SD_{intellig} = \sum_i I_i SD_i \tag{19}$$

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

$$SD_i = \frac{1}{(2^{1/6} - 2^{-1/6}) f_i^c} \int_{2^{-1/6} f_i^c}^{2^{1/6} f_i^c} |10 \log_{10} G^s(f)| df \tag{20}$$



(a) SNR improvement for variable SDW-MWF and different settings of μ



(b) Speech distortion for variable SDW-MWF and different settings of μ

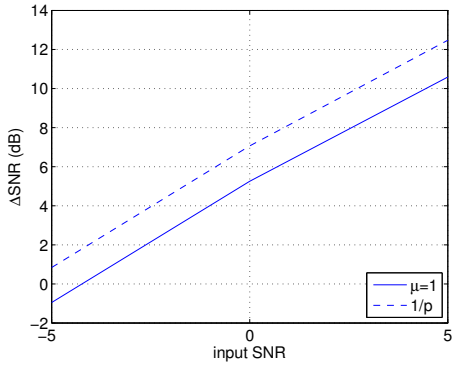
Fig. 1. A comparison of variable SDW-MWF with SDW-MWF with fixed settings of μ

with the center frequencies f_i^c and $G^s(f)$ the power spectral transfer function for the speech component from the input to the output of the noise reduction algorithm.

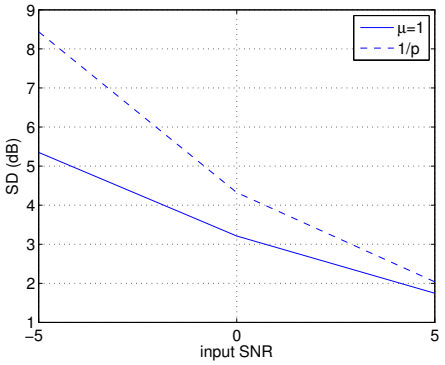
4.2. Variable vs. fixed SDW-MWF

In the first experiment the variable SDW-MWF is compared to SDW-MWF with different values of μ at input SNR 0dB. The SNR improvement is shown in figure 1(a). The SNR improvement for the SDW-MWF with different μ's are shown with the solid line and here the SNR improvement is as expected increasing with $\mu > 1$. On the other hand, the speech distortion is also increased which is shown in figure 1(b). The variable SDW-MWF shows that the SNR improvement is achieved at lower speech distortion. The reason for this is that the noise dominant segments are suppressed more compared to the speech dominant segments, resulting in an improved SNR at lower speech distortion.

In the second experiment the variable SDW-MWF is compared to SDW-MWF with $\mu = 1$ at input SNR -5dB to 5dB. The SNR improvement for different input SNR is shown in



(a) SNR improvement for variable SDW-MWF at different input SNR



(b) Speech distortion for variable SDW-MWF at different input SNR

Fig. 2. A comparison of variable SDW-MWF with SDW-MWF with $\mu = 1$ at different input SNR

figure 2(a). The solid line shows the SNR improvement for $\mu = 1$ which shows less SNR improvement compared to the variable SDW-MWF. As expected the speech distortion for $\mu = 1$ is still lower compared to the variable SDW-MWF at different input SNR. It is worth noting that at low input SNR like -5dB the SNR improvement comes at the cost of a much higher speech distortion. Whereas, at high input SNR e.g. 5dB the SNR improvement is achieved with a speech distortion close to the case with $\mu = 1$.

5. CONCLUSION

In this paper, we have presented a variable SDW-MWF that makes a trade-off between noise reduction and speech distortion based on the soft output VAD i.e. probability that speech is present in a given signal segment. Through simulations we have shown that with a variable SDW-MWF the noise reduction performance can be improved without increasing the speech distortion compared to the SDW-MWF with a fixed trade-off parameter.

6. REFERENCES

[1] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Communication*, vol. 7-8, pp. 636–656, July 2007.

[2] A. Spriet, M. Moonen, and J. Wouters, "Stochastic gradient based implementation of spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction in hearing aids," *IEEE Transactions on Signal Processing*, vol. 53, no. 3, pp. 911–625, Mar. 2005.

[3] R. J. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[4] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[5] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, Sept. 2003.

[6] K. Ngo, S. Doclo, A. Spriet, M. Moonen, J. wouters, and S.H. Jensen, "An integrated approach for noise reduction and dynamic range compression in hearing aids," *accepted for publication in Proc. 16th European Signal Processing Conference (EUSIPCO), Lausanne, Switzerland*, Aug. 2008.

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

[8] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function gsc and postfiltering," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

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

[10] Acoustical Society of America, "ANSI S3.5-1997 American National Standard Methods for calculation of the speech intelligibility index," June 1997.