A FLEXIBLE SPEECH DISTORTION WEIGHTED MULTI-CHANNEL WIENER FILTER FOR NOISE REDUCTION IN HEARING AIDS

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

[1]Katholieke Universiteit Leuven, ESAT-SCD, Leuven, Belgium
[2]Aalborg University, Dept. Electronic Systems, Aalborg, Denmark
[3]Katholieke Universiteit Leuven, ExpORL, O. & N2, Leuven, Belgium

ABSTRACT

In this paper, a multi-channel noise reduction algorithm is presented based on a Speech Distortion Weighted Multi-channel Wiener Filter (SDW-MWF) approach that incorporates a flexible weighting factor. A typical SDW-MWF uses a fixed weighting factor to trade-off between noise reduction and speech distortion without taking speech presence or speech absence into account. Consequently, the improvement in noise reduction comes at the cost of a higher speech distortion since the speech dominant segments and the noise dominant segments are weighted equally. Based on a two-state speech model with a noise-only and a speech+noise state, a solution is introduced that allows for a more flexible trade-off between noise reduction and speech distortion. Experimental results with hearing aid scenarios demonstrate that the proposed SDW-MWF incorporating the flexible weighting factor improves the signal-to-noise-ratio with lower speech distortion compared to a typical SDW-MWF and the SDW-MWF incorporating the conditional speech presence probability (SPP).

Index Terms— Multi-channel Wiener filter, noise reduction, distortion, speech presence probability, hearing aids.

1. INTRODUCTION

Background noise (from competing speakers, traffic etc.) is a significant problem for hearing impaired people who indeed have more difficulty understanding speech in noise and so in general need a higher signal-to-noise-ratio (SNR) than people with normal hearing [1]. The objective of these noise reduction algorithms is to maximally reduce the noise while minimizing speech distortion. In most scenarios, the desired speaker and the noise sources are physically located at different positions. Multi-channel noise reduction algorithms can then exploit both spectral and spatial characteristics of the speech and the noise. Another known multi-channel noise reduction algorithm is the Speech Distortion Weighted MWF (SDW-MWF) that provides an MMSE estimate of the speech component in one of the input signals [2][3].

Traditionally, these multi-channel noise reduction algorithms adopt a (short-time) fixed filtering under the implicit hypothesis that the speech is present at all time. However, while the noise can indeed be continously present, the speech signal typically contains many pauses. Furthermore, the speech may not be present at all frequencies even during speech segments. It has been shown in single-channel noise reduction algorithms that by incorporating the conditional SPP in the gain function or in the noise spectrum estimation a better performance can be achieved compared to traditional methods [4][5]. A typical SDW-MWF uses a fixed weighting factor to trade-off between noise reduction and speech distortion without taking speech presence or speech absence into account. This means that the speech dominant segments and the noise dominant segments are weighted equally in the noise reduction process. Consequently, the improvement in noise reduction comes at the cost of a higher speech distortion. In [6][7] an SDW-MWF approach that incorporates the conditional SPP in the trade-off between noise reduction and speech distortion has been introduced. In speech dominant segments it is then desirable to have less noise reduction to avoid speech distortion, while in noise dominant segments it is desirable to have as much noise reduction as possible.

This paper presents an SDW-MWF approach that incorporates a flexible weighting factor based on a two-state speech model with a noise-only and a speech+noise state. The flexible weighting factor is introduced to allow for a more flexible trade-off between noise reduction and speech distortion. Experimental results with hearing aid scenarios demonstrate that the proposed SDW-MWF incorporating a flexible weighting factor improves the signal-to-noise-ratio with lower speech distortion compared to a typical SDW-MWF and the SDW-MWF incorporating the conditional SPP.

The paper is organised as follows. Section 2 describes the general set-up and the multi-channel Wiener filter. Section 3 explains the concept behind introducing the flexible weighting factor in the SDW-MWF. In Section 4 experimental results are presented. The work is summarized in Section 5.

2. MULTI-CHANNEL WIENER FILTER

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

$$X_i(k,l) = X_i^s(k,l) + X_i^n(k,l) \tag{1}$$

where k is the frequency bin index, and l the frame index of a short-time Fourier transform (STFT), and the superscripts s and n are used to refer to the speech and the noise contribution in a signal, respectively. Let $\mathbf{X}(k,l) \in \mathbb{C}^{M\times 1}$ be defined as the stacked vector

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

This research work was carried out at the ESAT Laboratory of Katholieke Universiteit Leuven, in the frame of the EST-SIGNAL Marie-Curie Fellowship program (http://est-signal.i3s.unice.fr) under contract No. MEST-CT-2005-021175, the Concerted Research Action GOA-MaNet, 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, and the K.U.Leuven Research Council CoE EF/05/006 Optimization in Engineering (OPTEC). The scientific responsibility is assumed by its authors.

where the superscript T denotes the transpose. The MWF optimally estimates the speech signal, based on a Minimum Mean Squared Error (MMSE) criterion, i.e.,

$$\mathbf{W}_{\text{MMSE}}(k,l) = \arg\min_{\mathbf{W}} \ \varepsilon\{|X_1^s(k,l) - \mathbf{W}^H\mathbf{X}(k,l)|^2\} \quad (4)$$

where $\varepsilon\{\}$ denotes the expectation operator, H denotes Hermitian transpose and the desired signal in this case is the (unknown) speech component $X_1^s(k,l)$ in the first microphone signal. The MWF has been extended to the SDW-MWF$_\mu$ that allows for a trade-off between noise reduction and speech distortion using a weighting factor μ [2][3]. If the speech and the noise signals are statistically independent the design criterion of the SDW-MWF$_\mu$ is given by

$$\mathbf{W}_\mu(k,l) = \arg\min_{\mathbf{W}} \ \varepsilon\{|X_1^s(k,l) - \mathbf{W}^H\mathbf{X}^s(k,l)|^2\} + \mu\varepsilon\{|\mathbf{W}^H\mathbf{X}^n(k,l)|^2\}. \quad (5)$$

The SDW-MWF$_\mu$ is then given by

$$\mathbf{W}_\mu(k,l) = \left[\mathbf{R}^s(k,l) + \mu\mathbf{R}^n(k,l)\right]^{-1}\mathbf{R}^s(k,l)\mathbf{e}_1 \quad (6)$$

where the $M\times 1$ vector $\mathbf{e}_1$ equals the first canonical vector defined as $\mathbf{e}_1 = [1 \quad 0 \quad ... \quad 0]^T$ and the correlation matrices can be estimated as

$$H_0(k,l): \begin{cases} \mathbf{R}^n(k,l) = \alpha_n\mathbf{R}_n(k,l) + (1-\alpha_n)\mathbf{X}(k,l)\mathbf{X}^H(k,l) \\ \mathbf{R}^x(k,l) = \mathbf{R}_x(k,l) \end{cases}$$
$$H_1(k,l): \begin{cases} \mathbf{R}^x(k,l) = \alpha_x\mathbf{R}^x(k,l) + (1-\alpha_x)\mathbf{X}(k,l)\mathbf{X}^H(k,l) \\ \mathbf{R}^n(k,l) = \mathbf{R}^n(k,l) \end{cases} \quad (7)$$

where $H_0(k,l)$ and $H_1(k,l)$ represent speech absence and speech presence events in frequency bin k and frame l, respectively. The second-order statistics of the noise are assumed to be (short-term) stationary which means that $\mathbf{R}^s(k,l)$ can be estimated as $\mathbf{R}^s(k,l) = \mathbf{R}^x(k,l) - \mathbf{R}^n(k,l)$. Looking at (7) it is clear that $\mathbf{R}_x(k,l)$ and $\mathbf{R}_n(k,l)$ are updated at different time instant based on $H_0(k,l)$ and $H_1(k,l)$. Furthermore, an averaging time window of 2-3s (defined by α_n and α_x) is typically used to achieve a reliable estimate. Another aspect is the μ in (6) which is a fixed value for each frame and each frequency. This puts a limitation of the tracking capabilities since speech and noise are non-stationary and can be considered stationary only in a short time window, e.g., 8-20ms [1].

2.1. SDW-MWF incorporating the conditional Speech Presence probability (SDW-MWF$_{\text{SPP}}$)

A two-state model for speech events can be expressed given two hypotheses $H_0(k,l)$ and $H_1(k,l)$ which represent speech absence and speech presence in frequency bin k and frame l, respectively, i.e.,

$$H_0(k,l): X_i(k,l) = X_i^n(k,l) + 0\cdot X_i^s(k,l)$$
$$H_1(k,l): X_i(k,l) = X_i^n(k,l) + 1\cdot X_i^s(k,l), \quad (8)$$

where the i-th microphone signal is used as a reference (in our case the first microphone signal $X_1(k,l)$ is used). The inclusion of the second term in the definition of H_0 will be explained in Section 3. The conditional SPP $p(k,l) \triangleq P(H_1(k,l)|X_i(k,l))$ can be written as [5]

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l))\exp(-\upsilon(k,l))\right\}^{-1} \quad (9)$$

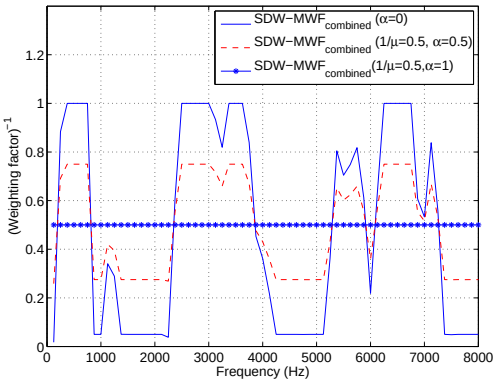


Fig. 1. Different configuration of (weighting factor)$^{-1}$.

where $q(k,l) \triangleq P(H_0(k,l))$ is the a priori speech absence probability (SAP), $\upsilon(k,l) \triangleq \frac{\gamma(k,l)\xi(k,l)}{(1+\xi(k,l))}$ such that $\xi(k,l)$ and $\gamma(k,l)$ denote the a priori SNR and the a posteriori SNR, respectively. Details on the estimation of the SAP, the a priori SNR and the a posteriori SNR can be found in [5][6].

For the sake of conciseness the frequency bin index k and frame index l are omitted from now on in $\mathbf{X}(k,l)$, $\mathbf{X}^s(k,l)$, $\mathbf{X}^n(k,l)$ and $X_1^s(k,l)$.

2.2. Derivation of SDW-MWF$_{\text{SPP}}$

The conditional SPP in (9) and the two-state model in (8) for speech events can be incorporated into the optimization criterion of the SDW-MWFμ, leading to a weighted average where the first term corresponds to H_1 and is weighted by the probability that speech is present, while the second term corresponds to H_0 and is weighted by the probability that speech is absent, i.e.,

$$\mathbf{W}_{\text{SPP}}(k,l) = \arg\min_{\mathbf{W}} \ p(k,l)\varepsilon\{|X_1^s - \mathbf{W}^H\mathbf{X}|^2|H_1\} + (1-p(k,l))\varepsilon\{|\mathbf{W}^H\mathbf{X}|^2|H_0\} \quad (10)$$

where $p(k,l)$ is the conditional probability that speech is present and $(1-p(k,l))$ is the conditional probability that speech is absent. The solution is then given by

$$\mathbf{W}_{\text{SPP}}(k,l) = \left[\mathbf{R}^s(k,l) + \left(\tfrac{1}{p(k,l)}\right)\mathbf{R}^n(k,l)\right]^{-1}\mathbf{R}^s(k,l)\mathbf{e}_1. \quad (11)$$

The SDW-MWF$_{\text{SPP}}$ offers more noise reduction when $p(k,l)$ is small, i.e., for noise dominant segments, and less noise reduction when $p(k,l)$ is large, i.e., for speech dominant segments making the SDW-MWF$_{\text{SPP}}$ change with a faster dynamic [6].

In [6] a combined solution SDW-MWF$_{\text{combined}}$ was also proposed, which in one extreme case corresponds to the SDW-MWF$_{\text{SPP}}$ and in the other extreme case corresponds to the SDW-MWF$_\mu$. Basically the term $\frac{1}{p(k,l)}$ is replaced with $\frac{1}{\alpha(\frac{1}{\mu})+(1-\alpha)p(k,l)}$ where α is a trade-off factor between SDW-MWF$_\mu$ and SDW-MWF$_{\text{SPP}}$. The (weighting factor)$^{-1}$ i.e. $\alpha(\frac{1}{\mu}) + (1-\alpha)p(k,l)$ is shown in Fig. 1 for different configurations. This clearly shows that the combined solution corresponds to a smoothing of the conditional SPP. Since the variations between the speech dominant segments and the noise dominant segments are reduced, the distortion is also reduced.

3. SDW-MWF INCORPORATING A FLEXIBLE WEIGHTING FACTOR (SDW-MWF$_{\text{FLEX}}$)

First, it is clear that the noise reduction in the H_0 state and the H_1 state have a different interpretation, i.e.,

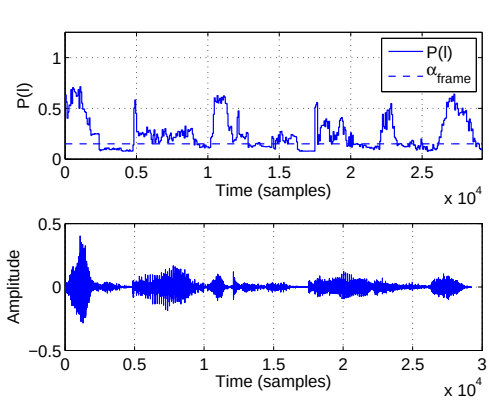


Fig. 2. Illustration of $P(l)$ for a given speech segment.

- Reducing the noise in the H_0 state can be related to increasing listening comfort, since speech is not present in the H_0 state, which means that a greater attenuation can be applied.
- Reducing the noise in the H_1 state is a more challenging task since this relates to speech intelligibility and hence the speech distortion weighted concept truly only makes sense in the H_1 state.

Secondly, as described in Section 2, the speech correlation matrix $\mathbf{R}^s(k,l)$ and the noise correlation matrix $\mathbf{R}^n(k,l)$ are estimated during H_1 and H_0, respectively. This means that, in theory the SDW-MWF could be an all zero vector during noise-only periods since then $\mathbf{R}^s(k,l) = 0$. In practice $\mathbf{R}^s(k,l)$ is "frozen" during noise-only periods where $\mathbf{R}^n(k,l)$ is updated. In fact this is in line with the definition of H_0 in (8), where the "0" indicate, that the speech X_i^s can have a non-zero $\mathbf{R}^s(k,l)$ in H_0, but is not transmitted into X_i. We then suggest, that if the H_0 state and the H_1 state can be properly detected a more flexible trade-off between noise reduction and speech distortion can be achieved. To this aim, the parameter $P(l)$ is introduced, which is a binary decision, obtained by averaging the conditional SPP $p(k,l)$ over all frequency bins k

$$P(l) = \begin{cases} 1 & \text{if} \quad \frac{1}{K}\sum_{k=1}^{K} p(k,l) \geq \alpha_{\text{frame}} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where $P(l) = 1$ means the H_1 state is detected and $P(l) = 0$ means the H_0 state is detected, and α_{frame} is a detection threshold. This $P(l)$ will be used in the operation of SDW-MWF$_{\text{Flex}}$. In Fig. 2 $P(l)$ is plotted for a given speech segment which shows that even in H_1 state there are some frames/samples where the conditional SPP is low. Notice that in this case the noise correlation matrice is kept fixed whereas $p(k,l)$ and $P(l)$ are continously updated. The two key ingredients of the proposed SDW-MWF$_{\text{Flex}}$ are now as follows:

- A weighting factor μ_{H_1} is introduced, which is a function of $p(k,l)$, that defines the amount of noise reduction that can be applied in the H_1 state.
- A weighting factor μ_{H_0} is introduced, which is a constant weighting factor, that defines the amount of noise reduction that can be applied in the H_0 state.

The SDW-MWF$_{\text{Flex}}$ weighting strategy is illustrated in Fig. 3 which shows the weighting factor as a function of $p(k,l)$. Notice that μ_{H_1} is defined here as $\min(\frac{1}{p(k,l)}, \alpha_{H_1})$, i.e., a function of the conditional SPP $\frac{1}{p(k,l)}$ and a lower threshold α_{H_1} which is introduced

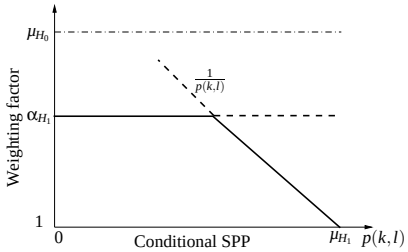


Fig. 3. The weighting factor used in SDW-MWF$_{\text{Flex}}$.

since speech may not be present in all frequency bins even in state H_1. The optimization criterion for SDW-MWF$_{\text{Flex}}$ is given by

$$\begin{aligned} \mathbf{W}_{\text{Flex}}(k,l) &= \arg\min_{\mathbf{W}} \\ & P(l)\Big[\max(p(k,l), \tfrac{1}{\alpha_{H_1}})\varepsilon\{|X_1^s - \mathbf{W}^H\mathbf{X}|^2|H_1\} \\ & + (1 - \max(p(k,l), \tfrac{1}{\alpha_{H_1}}))\varepsilon\{|\mathbf{W}^H\mathbf{X}|^2|H_0\}\Big] + \\ & (1-P(l))\Big[\tfrac{1}{\mu_{H_0}}\varepsilon\{|X_1^s - \mathbf{W}^H\mathbf{X}^s|^2\} + \varepsilon\{|\mathbf{W}^H\mathbf{X}^n|^2\}\Big] \\ &= \arg\min_{\mathbf{W}} \\ & \Big[P(l)\max(p(k,l), \tfrac{1}{\alpha_{H_1}}) + (1-P(l))\tfrac{1}{\mu_{H_0}}\Big] \\ & \varepsilon\{|X_1^s - \mathbf{W}^H\mathbf{X}^s|^2\} + \varepsilon\{|\mathbf{W}^H\mathbf{X}^n|^2\} \end{aligned} \tag{13}$$

The solution is given by

$$\mathbf{W}_{\text{Flex}}(k,l) = \Big[\mathbf{R}^s + \gamma(k,l)\mathbf{R}^n\Big]^{-1}\mathbf{R}^s\mathbf{e}_1 \tag{14}$$

with the weighting factor defined as

$$\begin{aligned} \gamma(k,l) &= \Big[P(l)\max(p(k,l), \tfrac{1}{\alpha_{H_1}}) + (1-P(l))\tfrac{1}{\mu_{H_0}}\Big]^{-1} \\ &= \Big[P(l)\min(\tfrac{1}{p(k,l)}, \alpha_{H_1}) + (1-P(l))\mu_{H_0}\Big]. \end{aligned} \tag{15}$$

4. EXPERIMENTAL RESULTS

In this section, experimental results for the proposed SDW-MWF$_{\text{Flex}}$ are presented and compared to SDW-MWF$_{\text{SPP}}$ and SDW-MWF$_{\mu}$.

4.1. Experimental set-up and performance measures

Simulations have been performed with a 2-microphone behind-the-ear hearing aid mounted on a CORTEX MK2 manikin. The loudspeakers (FOSTEX 6301B) are positioned at 1 meter from the center of the head. The reverberation time T_{60}=0.21s. The speech is located at 0° and the two multi-talker babble noise sources are located at 120° and 180°. The speech signal consists of male sentences from Hearing in Noise Test (HINT) for the measurement of speech reception thresholds in quiet and in noise and the noise signal consists of a multi-talker babble from Auditory Tests (Revised), Compact Disc, Auditec. The signals are sampled at 16kHz. An FFT length of 128 with 50% overlap was used. The parameters for estimating the conditional SPP are similiar as in [6].

To assess the noise reduction performance the intelligibility-weighted signal-to-noise ratio (SNR) [8] is used which is defined as

$$\Delta\text{SNR}_{\text{intellig}} = \sum_i I_i(\text{SNR}_{i,out} - \text{SNR}_{i,in}) \tag{16}$$

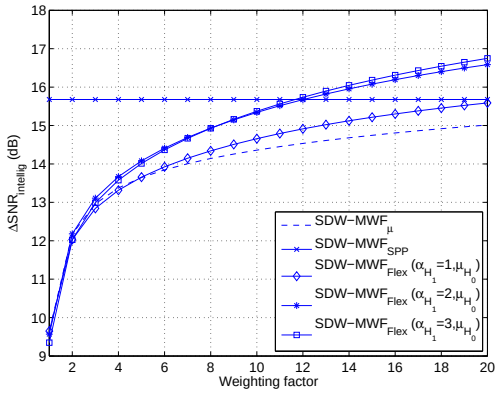


Fig. 4. SNR improvement for α_{H_1}=1,2 and 3 with variable μ_{H_0} compared to SDW-MWF$_\mu$ and SDW-MWF$_{\text{SPP}}$.

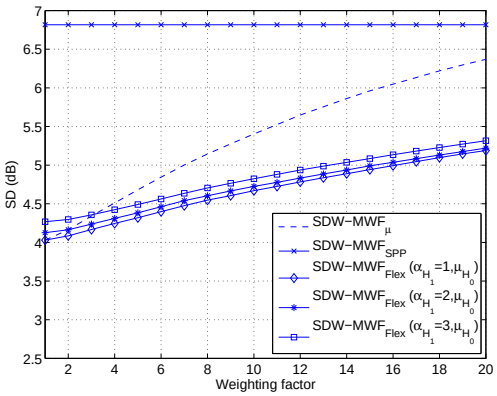


Fig. 5. Speech distortion for α_{H_1}=1,2 and 3 with variable μ_{H_0} compared to SDW-MWF$_\mu$ and SDW-MWF$_{\text{SPP}}$.

where I_i is the band importance function defined in ANSI S3.5-1997 [9] and where SNR$_{i,out}$ and SNR$_{i,in}$ represent the output SNR and the input SNR (in dB) of the *i*-th band, respectively. For measuring the signal distortion a frequency-weighted log-spectral signal distortion (SD) is used defined as

$$\text{SD} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\int_{f_l}^{f_u} w_{ERB}(f)\Big(10log_{10}\frac{P^s_{out,k}(f)}{P^s_{in,k}(f)}\Big)^2 df} \qquad (17)$$

where K is the number of frames, $P^s_{out,k}(f)$ is the output power spectrum of the kth frame, $P^s_{in,k}(f)$ is the input power spectrum of kth frame and f is the frequency index. The SD measure is calculated with a frequency-weighting factor $w_{ERB}(f)$ giving equal weight for each auditory critical band, as defined by the equivalent rectangular bandwidth (ERB) of the auditory filter [10]. Notice that the intelligibility-weighted SNR and the spectral distortion are only computed during frames of speech+noise.

4.2. Results

In this experiment, for the SDW-MWF$_{\text{Flex}}$, the α_{H_1} is fixed to 1,2 and 3, μ_{H_0} is increased from 1 to 20 and the conditional SPP $p(k,l)$ is estimated according to (9). For SDW-MWF$_\mu$, μ is increased from 1 to 20. The SNR improvement is shown in Fig. 4 and the speech distortion is shown Fig. 5. This shows, that the SDW-MWF$_{\text{Flex}}$ outperforms the SDW-MWF$_\mu$ and SDW-MWF$_{\text{SPP}}$ both in SNR improvement and in terms of speech distortion, when the weighting factor μ_{H_0} is increased. Increasing α_{H_1} does show a further improvement in SNR using SDW-MWF$_{\text{Flex}}$ with a small increase in speech distortion.

5. CONCLUSION

In this paper a noise reduction procedure SDW-MWF$_{\text{Flex}}$ has been presented that incorporates a flexible weighting factor to trade-off between noise reduction and speech distortion, which is an extension of the SDW-MWF$_{\text{SPP}}$ incorporating the conditional SPP. Based on a two-state speech model, with a noise-only (H_0) and a speech+noise (H_1) state, the goal of the SDW-MWF$_{\text{Flex}}$ is to apply an equal amount of noise reduction as in a typical SDW-MWF$_\mu$ in the H_0 state, while in the H_1 state, the goal is to preserve the speech by exploiting the conditional SPP. The SDW-MWF$_{\text{Flex}}$ is found to significantly improve the SNR while the speech distortion is kept low compared to SDW-MWF$_\mu$ and SDW-MWF$_{\text{SPP}}$.

6. REFERENCES

[1] H. Dillon, *Hearing Aids*, Boomerang Press, Turramurra, Australia, 2001.

[2] 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.

[3] 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.

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

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

[6] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Incorporating the conditional speech presence probability in multi-channel wiener filter based noise reduction in hearing aids," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 930625, 11 pages, 2009.

[7] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S.H. Jensen, "Variable speech distortion weighted multichannel wiener filter based on soft output voice activity detection for noise reduction in hearing aids," in *Proc. 11th IWAENC*, Seattle, USA, 2008.

[8] 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.

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

[10] B Moore, *An Introduction to the Psychology of Hearing*, Academic Press, 5th ed edition, 2003.