Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 930625, 11 pages
doi:10.1155/2009/930625

Research Article

Incorporating the Conditional Speech Presence Probability in Multi-Channel Wiener Filter Based Noise Reduction in Hearing Aids

Kim Ngo (EURASIP Member),[1] Ann Spriet,[1,2] Marc Moonen (EURASIP Member),[1] Jan Wouters,[2] and Søren Holdt Jensen (EURASIP Member)[3]

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

Correspondence should be addressed to Kim Ngo, kim.ngo@esat.kuleuven.be

Received 15 December 2008; Revised 30 March 2009; Accepted 2 June 2009

Recommended by Walter Kellermann

A multi-channel noise reduction technique is presented based on a Speech Distortion-Weighted Multi-channel Wiener Filter (SDW-MWF) approach that incorporates the conditional Speech Presence Probability (SPP). A traditional SDW-MWF uses a fixed parameter to a trade-off between noise reduction and speech distortion without taking speech presence 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. Incorporating the conditional SPP in SDW-MWF allows to exploit the fact that speech may not be present at all frequencies and at all times, while the noise can indeed be continuously present. 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. Experimental results with hearing aid scenarios demonstrate that the proposed SDW-MWF incorporating the conditional SPP improves the signal-to-noise ratio compared to a traditional SDW-MWF.

1. Introduction

Background noise (multiple speakers, traffic, etc.) is a significant problem for hearing aid users and is especially damaging to speech intelligibility. Hearing-impaired people have more difficulty understanding speech in noise and in general need a higher signal-to-noiseratio (SNR) than people with normal hearing to communicate effectively [1]. To overcome this problem both single-channel and multi-channel noise reduction algorithms have been proposed. The objective of these noise reduction algorithms is to maximally reduce the noise while minimizing speech distortion.

One of the first proposed single-channel noise reduction algorithms is spectral subtraction [2], which is based on the assumption that the noise is additive, and the clean speech spectrum can be obtained by subtracting an estimate of the noise spectrum from the noisy speech spectrum. The noise spectrum is updated during periods where the speech is absent, as detected by a Voice Activity Detection (VAD). Another well-known single-channel noise reduction technique is the Ephraim and Malah noise suppressor [3, 4], which estimates the amplitude of the clean speech spectrum in the spectral or in the log-spectral domain based on a Minimum Mean Square Error (MMSE) criterion. Common for these techniques are usually noticeable artifacts known as musical noise [5] mainly caused by the short-time spectral attenuation, the nonlinear filtering and, an inaccurate estimate of the noise characteristic. A limitation of single-channel noise reduction is that only differences in temporal and spectral signal characteristics can be exploited. In a multiple speaker scenario also known as the cocktail party problems the speech and the noise considerably

overlap in time and frequency. This makes it difficult for single-channel noise reduction schemes to suppress the noise without reducing speech intelligibility and introducing speech distortion or musical noise.

However, in most scenarios, the desired speaker and the disturbing noise sources are physically located at different positions. Multi-channel noise reduction can then exploit the spatial diversity, that is, exploit both spectral and spatial characteristics of the speech and the noise sources. The Frost beamformer and the Generalized Sidelobe Canceler [6–8] are well-known multi-channel noise reduction techniques. The basic idea is to steer a beam toward the desired speaker while reducing the background noise coming from other directions. Another known multi-channel noise reduction technique is the Multi-channel Wiener filter (MWF) that provides an MMSE estimate of the speech component in one of the microphone signals. The extension from MWF to Speech Distortion-Weighted MWF (SDW-MWF) [9, 10] allows for a trade-off between noise reduction and speech distortion.

Traditionally, these multi-channel noise reduction algorithms adopt a (short-time) fixed filtering under the implicit hypothesis that the clean speech is present at all time. However, the speech signal typically contains many pauses while the noise can indeed be continuously present. Furthermore, the speech may not be present at all frequencies even during voiced speech segments. It has been shown in single-channel noise reduction schemes that incorporating the conditional Speech Presence Probability (SPP) in the gain function or in the noise spectrum estimation better performance can be achieved compared to traditional methods [4, 11–13]. In these approaches the conditional SPP is estimated for each frequency bin and each frame by a soft-decision approach, which exploits the strong correlation of speech presence in neighboring frequency bins of consecutive frames.

A traditional SDW-MWF uses a fixed parameter to a 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. A variable SDW-MWF was introduced in [14] based on soft output voice activity detection to a trade-off between speech dominant segments and noise dominant segments. This paper presents an SDW-MWF approach that incorporates the conditional SPP in the trade-off between noise reduction and speech distortion. 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. Furthermore, a combined solution is introduced that in one extreme case corresponds to an SDW-MWF incorporating the conditional SPP and in the other extreme case corresponds to a traditional SDW-MWF solution. Experimental results with hearing aid scenarios demonstrate that the proposed SDW-MWF incorporating the conditional SPP improves the SNR compared to a traditional SDW-MWF.

The paper is organized as follows. Section 2 describes the system model and the general set-up of a multi-channel noise reduction algorithm. The motivation is given in Section 3. Section 4 explains the estimation of the conditional SPP. Section 5 explains the derivation of the SDW-MWF incorporating the conditional SPP. In Section 6 experimental results are presented. The work is summarized in Section 7.

2. System Model

A general set-up of a multi-channel noise reduction is shown in Figure 1 with M microphones in an environment with one or more noise sources and a desired speaker. Let $X_i(k,l), i = 1,\ldots,M$, denote the 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, l the frame index, 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

$$\begin{aligned}\mathbf{X}(k,l) &= [X_1(k,l)\;\; X_2(k,l)\cdots X_M(k,l)]^T \\ &= \mathbf{X}^s(k,l) + \mathbf{X}^n(k,l),\end{aligned} \tag{2}$$

where the superscript T denotes the transpose. In addition, we define the noise and the speech correlation matrices as

$$\begin{aligned}\mathbf{R}^n(k,l) &= \varepsilon\left\{\mathbf{X}^n(k,l)\mathbf{X}^{n,H}(k,l)\right\}, \\ \mathbf{R}^s(k,l) &= \varepsilon\left\{\mathbf{X}^s(k,l)\mathbf{X}^{s,H}(k,l)\right\},\end{aligned} \tag{3}$$

where $\varepsilon\{\}$ denotes the expectation operator, and H denotes Hermitian transpose.

2.1. Multi-channel Wiener Filter (MWF and SDW-MWF). The MWF optimally estimates a desired signal, based on a Minimum Mean Squared Error (MMSE) criterion, that is,

$$\mathbf{W}^*(k,l) = \arg\min_{\mathbf{W}} \varepsilon\left\{\left|X_1^s(k,l) - \mathbf{W}^H\mathbf{X}(k,l)\right|^2\right\}, \tag{4}$$

where the desired signal in this case is the speech component $X_1^s(k,l)$ 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 [9, 10]. The design criterion of the SDW-MWF is given by

$$\begin{aligned}\mathbf{W}^*(k,l) = \arg\min_{\mathbf{W}} &\varepsilon\left\{\left|X_1^s(k,l) - \mathbf{W}^H\mathbf{X}^s(k,l)\right|^2\right\} \\ &+ \mu\varepsilon\left\{\left|\mathbf{W}^H\mathbf{X}^n(k,l)\right|^2\right\}.\end{aligned} \tag{5}$$

If 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

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

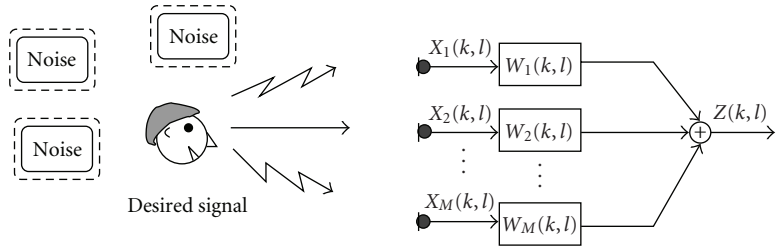

Figure 1: Multi-channel noise reduction set-up in an environment with one or more noise sources and a desired speaker.

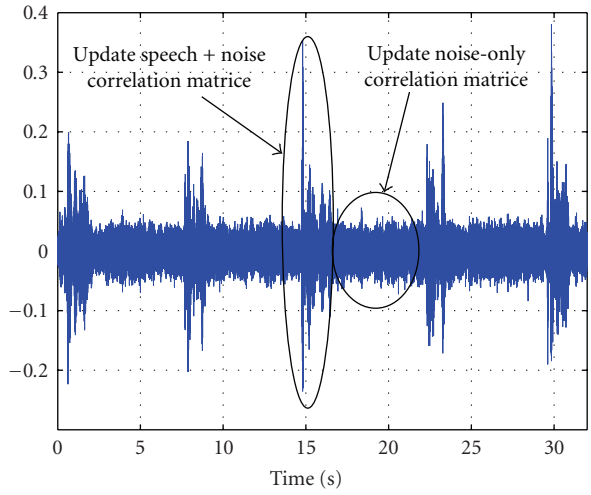

Figure 2: Illustration of a concatenated noisy speech signal with noise-only periods which is a typical input signal for multimicrophone noise reduction.

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

$$Z(k,l) = \mathbf{W}^{*,H}(k,l)\mathbf{X}(k,l). \tag{7}$$

2.2. MWF in Practice. A typical input signal for a multi-channel noise reduction is shown in Figure 2, where several speech sentences are concatenated with sufficient noise-only periods. By using a VAD the speech+noise and noise-only periods can be detected, and the corresponding correlation matrices can be estimated/updated. MWF is uniquely based on the second-order statistics, and in the estimation of the speech+noise and the noise-only correlation matrices an averaging time window of 2-3 seconds is typically used to achieve a reliable estimate. This suggests that the noise reduction performance of the MWF depends on the long-term average of the spectral and the spatial characteristics of the speech and the noise sources. In practice, this means that the MWF can only work well if the long-term spectral and/or spatial characteristics of the speech and the noise are slowly time-varying.

3. Motivation

The success of any NR algorithm is based on how much information is available about the speech and the noise [1, 15, 16]. In general speech and noise can be nonstationary both temporally, spectrally, and spatially. Speech is a spectrally nonstationary signal and can be considered stationary only in a short time window of 20–30 miliseconds. Background noise such as multitalker babble is also considered to be spectrally non-stationary. Furthermore, the speech characteristic contains many pauses while the noise can be continuously present. These properties are usually not taken into consideration in multi-channel noise reduction algorithms since the spatial characteristics are assumed to be more or less stationary, which then indeed justifies the long-term averaging of the correlation matrices. This long-term averaging basically eliminates any short-time effects, such as musical noise, that typically occur in single-channel noise reduction.

The motivation behind introducing the conditional SPP in SDW-MWF is to allow for a faster tracking of the non-stationarity of the speech and the noise as well as for exploiting the fact that speech may not be present at all time. This then allows to apply a different weight to speech dominant segments and to noise dominant segments in the noise reduction process. Furthermore incorporating the conditional SPP in the SDW-MWF also allows the NR to be applied in a narrow frequency band since the conditional SPP is estimated for each frequency bin; see Section 4.

4. Speech Presence Probability Estimation

The conditional SPP is estimated for each frequency bin and each frame by a soft-decision approach [12, 15, 17], which exploits the strong correlation of speech presence in neighboring frequency bins of consecutive frames.

4.1. Two-State Speech Model. A two-state model for speech events can be expressed given by two hypotheses $H_0(k,l)$

and $H_1(k,l)$ which represent speech absence and speech presence in each frequency bin, respectively, that is,

$$\begin{aligned} H_0(k,l) &: X_i(k,l) = X_i^n(k,l), \\ H_1(k,l) &: X_i(k,l) = X_i^s(k,l) + X_i^n(k,l). \end{aligned} \tag{8}$$

Assuming a complex Gaussian distribution of the Short-Time Fourier Transform (STFT) coefficients for both the speech and the noise, the conditional Probability Density Functions (PDFs) of the observed signals are given by

$$\begin{aligned} p(X_i(k,l) \mid H_0(k,l)) &= \frac{1}{\pi\lambda_i^n(k,l)} \exp\left\{-\frac{|X_i(k,l)|^2}{\lambda_i^n(k,l)}\right\}, \\ p(X_i(k,l) \mid H_1(k,l)) &= \frac{1}{\pi(\lambda_i^s(k,l)+\lambda_i^n(k,l))} \\ &\quad \times \exp\left\{-\frac{|X_i(k,l)|^2}{\lambda_i^s(k,l)+\lambda_i^n(k,l)}\right\}, \end{aligned} \tag{9}$$

where $\lambda_i^s(k,l) \triangleq \varepsilon\{|X_i^s(k,l)|^2 H_1(k,l)\}$ and $\lambda_i^n \triangleq \varepsilon\{|X_i^n(k,l)|^2\}$ denote the power spectrum of the speech and the noise, respectively. Applying Bayes rule, the conditional SPP $p(k,l) \triangleq P(H_1(k,l) \mid X_i(k,l))$ can be written as [4]

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l))\exp(-\upsilon(k,l))\right\}^{-1}, \tag{10}$$

where $q(k,l) \triangleq P(H_0(k,l))$ is the a priori Speech Absence Probability (SAP); $\xi(k,l)$ and $\gamma(k,l)$ denote the a priori SNR and a posteriori SNR, respectively,

$$\begin{aligned} \xi(k,l) &\triangleq \frac{\lambda_i^s(k,l)}{\lambda_i^n(k,l)}, \qquad \gamma(k,l) \triangleq \frac{|X_i(k,l)|^2}{\lambda_i^n(k,l)}, \\ \upsilon(k,l) &\triangleq \frac{\gamma(k,l)\xi(k,l)}{(1+\xi(k,l))}. \end{aligned} \tag{11}$$

The noise power spectrum λ_i^n is estimated using recursive averaging during periods where the speech is absence, that is,

$$\begin{aligned} H_0'(k,l) &: \hat{\lambda}_i^n(k,l+1) = \rho\hat{\lambda}_i^n(k,l) + (1-\rho)|X_i(k,l)|^2, \\ H_1'(k,l) &: \hat{\lambda}_i^n(k,l+1) = \hat{\lambda}_i^n(k,l). \end{aligned} \tag{12}$$

where ρ is an averaging parameter, and $H_0'(k,l)$ and $H_1'(k,l)$ represents speech absence and speech presence, respectively. The noise power spectrum is updated using a perfect VAD such that the noise power is updated at the same time as the noise correlation matrix; see Figure 2. The noise spectrum can also be estimated by using the Minima Controlled Recursive Averaging approach presented here [13]. The main issue in estimating the conditional SPP $p(k,l)$ is to have reliable estimates of the a priori SNR and the a priori SAP used in (10). Since speech has a non-stationary characteristic the a priori SNR and the a priori SAP are estimated for each frequency bin of the noisy speech.

4.2. A Priori SNR Estimation. The decision-directed approach of Ephraim and Malah [4, 12, 17] is widely used for estimating the a priori SNR and is given by

$$\hat{\xi}(k,l) = \kappa\frac{|\hat{X}_i(k,l-1)|^2}{\lambda_i^n(k,l-1)} + (1-\kappa)\max\{\gamma(k,l)-1, 0\}, \tag{13}$$

where $|\hat{X}_i(k,l-1)|^2$ represents an estimate of the clean speech spectrum, and κ is a weighting factor that controls the trade-off between noise reduction and speech distortion [4, 5]. The first term corresponds to the SNR from the previous enhanced frame, and the second term is the estimated SNR for the current frame.

4.3. A Priori SAP Estimation. Reliable estimation of the a priori SNR is important since it is used in the estimation for the a priori SAP. In [12, 17] an a priori SAP estimator is proposed based on the time-frequency distribution of the estimated a priori SNR $\hat{\xi}(k,l)$. The estimation is based on three parameters that each exploits the strong correlation of speech presence in neighboring frequency bins of consecutive frames. The first step is to apply a recursive averaging to the a priori SNR, that is,

$$\zeta(k,l) = \beta\zeta(k,l-1) + (1-\beta)\hat{\xi}(k,l-1), \tag{14}$$

where β is the averaging parameter. In the second step a global and local averaging is applied to $\zeta(k,l)$ in the frequency domain. Local means that the a priori SNR is avaraged over a small number of frequency bins (small bandwidth), and global means that the a priori SNR is averaged over a larger number of frequency bins (larger bandwidth). The local and global averaging of the a priori SNR is given by

$$\zeta_\eta(k,l) = \sum_{i=-\omega_\eta}^{i=\omega_\eta} h_\eta(i)\zeta(k-i,l), \tag{15}$$

where the subscript η represents either local or global averaging, and h_η is a normalized Hanning window of size $2\omega_\eta + 1$. The local and global averaging of the a priori SNR is then normalized to values between 0 and 1 before it is mapped into the following threshold function:

$$P_\eta(k,l) = \begin{cases} 0, & \text{if } \zeta_\eta(k,l) \le \zeta_{\min} \\ 1, & \text{if } \zeta_\eta(k,l) \ge \zeta_{\max} \\ \dfrac{\log\left(\zeta_\eta(k,l)/\zeta_{\min}\right)}{\log(\zeta_{\max}/\zeta_{\min})}, & \text{otherwise} \end{cases} \tag{16}$$

where $P_{\text{local}}(k,l)$ is the likelihood of speech presence when the a priori SNR is avaraged over a small number of frequency bins, and $P_{\text{global}}(k,l)$ is the likelihood of speech presence when the a priori SNR is averaged over a larger number of frequency bins. $\zeta_{\min}$ and $\zeta_{\max}$ are empirical constants that decide the threshold for speech or noise. The last term

$P_{\text{frame}}(l)$ represents the likelihood of speech presence in a given frame based on the a priori SNR averaged over all frequency bins, that is,

$$\zeta_{\text{frame}}(l) = \underset{1\le k\le N/2+1}{\text{mean}}\{\zeta(k,l)\}, \tag{17}$$

where N is the FFT-size. A pseudocode for the computation of $P_{\text{frame}}(l)$ is given by

if $\zeta_{\text{frame}}(l) > \zeta_{\min}$ **then**

  if $\zeta_{\text{frame}}(l) > \zeta_{\text{frame}}(l-1)$ **then**

    $P_{\text{frame}}(l) = 1$

    $\zeta_{\text{peak}}(l) = \min\left\{\max\left[\zeta_{\text{frame}}(l), \zeta_{p\ \min}\right], \zeta_{p\ \ \max}\right\}$

  else (18)

    $P_{\text{frame}}(l) = \delta(l)$

  else

    $P_{\text{frame}}(l) = 0$

  end if

end if

where

$$\delta(l) = \begin{cases} 0, & \text{if } \zeta_{\text{frame}}(l) \le \zeta_{\text{peak}}(l)\cdot\zeta_{\min} \\ 1, & \text{if } \zeta_{\text{frame}}(l) \ge \zeta_{\text{peak}}(l)\cdot\zeta_{\max} \\ \dfrac{\log\left(\zeta_{\text{frame}}(l)/\zeta_{\text{peak}}(l)/\zeta_{\min}\right)}{\log(\zeta_{\max}/\zeta_{\min})}, & \text{otherwise} \end{cases} \tag{19}$$

represents a soft transition from speech noise, ζ_{peak} is a confined peak value of ζ_{frame}, and $\zeta_{p\ \min}$ and $\zeta_{p\ \max}$ are empirical constants that determine the delay of the transition. The proposed a priori SAP estimation is then obtained by

$$\hat{q}(k,l) = 1 - P_{\text{local}}(k,l)\cdot P_{\text{global}}(k,l)\cdot P_{\text{frame}}(l). \tag{20}$$

This means that if either of the previous frames or recent frequency bins does not contain speech, that is, if the three likelihood terms are small, then $\hat{q}(k,l)$ becomes larger and the conditional SPP $p(k,l)$ in (10) becomes smaller.

Two examples of the normalized a priori SNR for different frames are shown in Figures 3 and 4. If the lower threshold $\zeta_{\min}$ is set too high, then there is a greater chance for noise classification, and at the same time weaker frequency components might also be ignored. If $\zeta_{\min}$ in Figure 3 is increased, then the weak high-frequency component will be classified as noise. On the other hand if $\zeta_{\max}$ is increased in Figure 4, the weaker low-frequency component will not be classified as a speech dominant segment. The estimated conditional SPPs for the two examples given above are shown in Figures 5 and 6. As mentioned above the weak high-frequency component in Figure 5 will be ignored if $\zeta_{\min}$ is increased, and the speech dominant segment at low-frequency in Figure 6 will not be as significant if $\zeta_{\max}$ is increased. In general, classifying noise when speech is present is more harmful than classifying speech when noise is present. By setting $\zeta_{\min}$ and $\zeta_{\max}$ low more speech will be detected, and the same goes for the setting of $\zeta_{p\ \min}$ and $\zeta_{p\ \max}$. The goal is then to incorporate this conditional SPP into the SDW-MWF such that speech dominant segments will be attenuated less compared to speech dominant segments. By exploiting the conditional SPP shown in Figures 5 and 6 the noise can be reduced in a narrow frequency band, that is, when the conditional SPP is low.

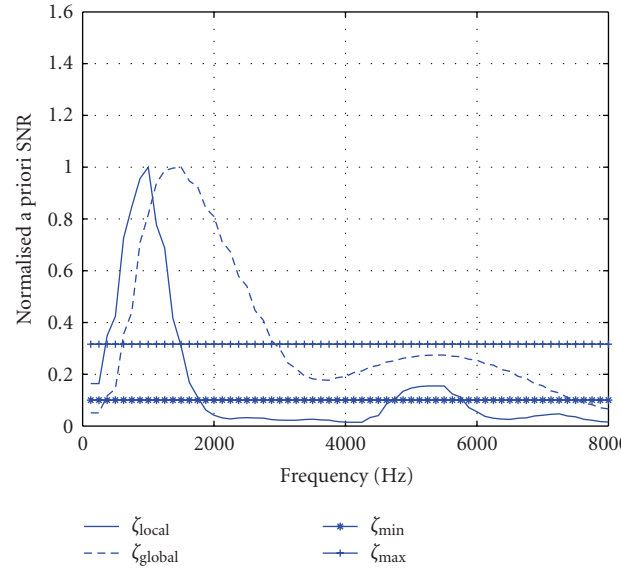

Figure 3: Local and global averaging of the a priori SNR for a given frame. Example of a high a priori SNR at low frequency.

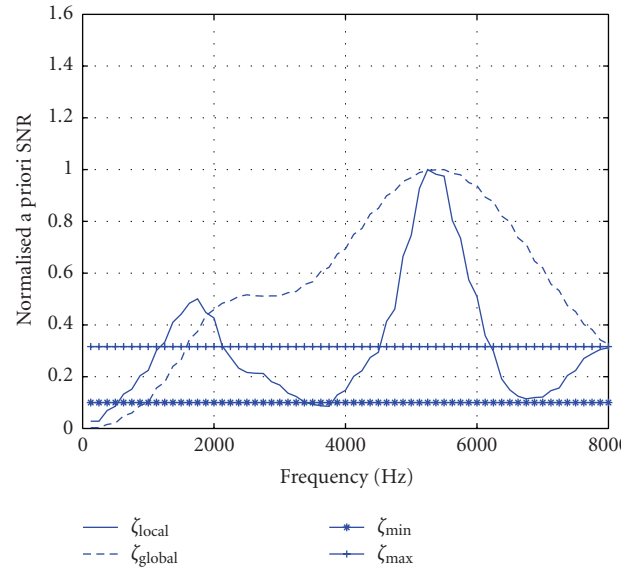

Figure 4: Local and global averaging of the a priori SNR for a given frame. Example of a high a priori SNR at high frequency.

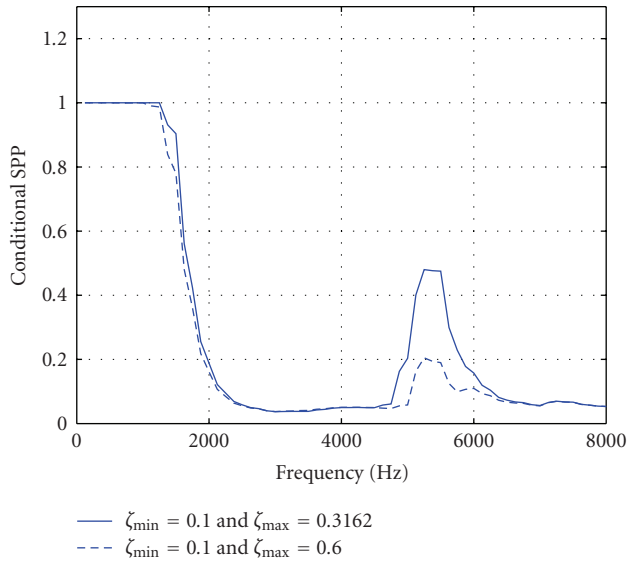

Figure 5: Conditional SPP with high-speech presence at low frequency.

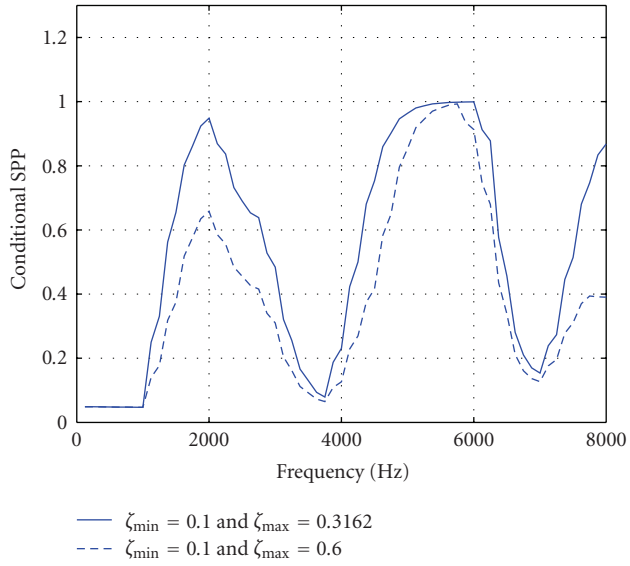

Figure 6: Conditional SPP with two distinct speech dominant segments.

The frequency bin index k and frame index l are omitted in the sequel for the sake of conciseness.

5. SDW-MWF Incorporating the Conditional Speech Presence Probability

In this section, we derive a modified SDW-MWF, which incorporates the conditional SPP in the filter estimation, which is referred to as SDW-MWF$_{\text{SPP}}$ from now on. Traditionally, the trade-off parameter in SDW-MWF$_\mu$ in (5) is set to a fixed value, and any 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 an SDW-MWF$_{SPP}$ it is possible to distinguish between the speech+noise segments and noise-only segments. The conditional SPP in (10) and the two-state model in (8) for speech events can be incorporated into the optimization criteria of the SDW-MWF, 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 that is weighted by the probability that speech is absent, that is,

$$\begin{aligned}\mathbf{W}^* = \arg\min_{\mathbf{W}} & p\varepsilon\left\{\left|X_1^s - \mathbf{W}^H\mathbf{X}\right|^2 \mid H_1\right\} \\ & + (1-p)\varepsilon\left\{\left|\mathbf{W}^H\mathbf{X}^n\right|^2\right\},\end{aligned} \tag{21}$$

where $p = P(H_1 \mid X_i)$ is the conditional probability that speech is present when observing X_i, and $(1-p) = P(H_0 \mid X_i)$ is the probability that speech is absent when observing X_i. The solution is then given by

$$\begin{aligned}\mathbf{W}^* &= \left(p\varepsilon\left\{\mathbf{X}\mathbf{X}^H \mid H_1\right\} + (1-p)\varepsilon\left\{\mathbf{X}^n\mathbf{X}^{n,H}\right\}\right)^{-1} \\ &\quad \times p\varepsilon\left\{\mathbf{X}^s X_1^{s,H} \mid H_1\right\} \\ &= \left(p\varepsilon\left\{\mathbf{X}^s\mathbf{X}^{s,H} \mid H_1\right\} + p\varepsilon\left\{\mathbf{X}^n\mathbf{X}^{n,H}\right\}\right. \\ &\quad \left. + (1-p)\varepsilon\left\{\mathbf{X}^n\mathbf{X}^{n,H}\right\}\right)^{-1} p\varepsilon\left\{\mathbf{X}^s X_1^{s,H} \mid H_1\right\} \\ &= \left(p\varepsilon\{\mathbf{X}^s\mathbf{X}^{s,H} \mid H_1\} + \varepsilon\{\mathbf{X}^n\mathbf{X}^{n,H}\}\right)^{-1} p\varepsilon\left\{\mathbf{X}^s X_1^{s,H} \mid H_1\right\}.\end{aligned} \tag{22}$$

The SDW-MWF incorporating the conditional SPP can then be written as

$$\mathbf{W}_{\text{SPP}}^* = \left(\mathbf{R}^s + \left(\frac{1}{p}\right)\mathbf{R}^n\right)^{-1}\mathbf{R}^s\mathbf{e}_1. \tag{23}$$

Compared to (6) with the fixed μ the term $1/p$, which is defined as the weighting factor, is now adjusted for each frequency bin and for each frame, making the SDW-MWF$_{\text{SPP}}$ changes with a faster dynamic. Figure 7 presents a block diagram of the proposed SDW-MWF$_{\text{SPP}}$. First an FFT is performed on each frame of the noisy speech. Then on the left-hand side the conditional SPP is estimated, which includes the estimation of the a posteriori SNR, the a priori SNR, and the a priori SAP. On the right-hand side the frequency domain correlation matrices are estimated, which are used to estimate the filter coefficients after weighting with the conditional SPP. Notice that the updates of the frequency domain correlation matrices are still based on a longer time window; see Section 2.2. The difference is now that the weights applied in the filter estimation are now changing for each frequency bin and each frame based on the conditional SPP. The last steps include the filtering operation and the IFFT. The conditional SPP weighting factor $1/p$ offers more

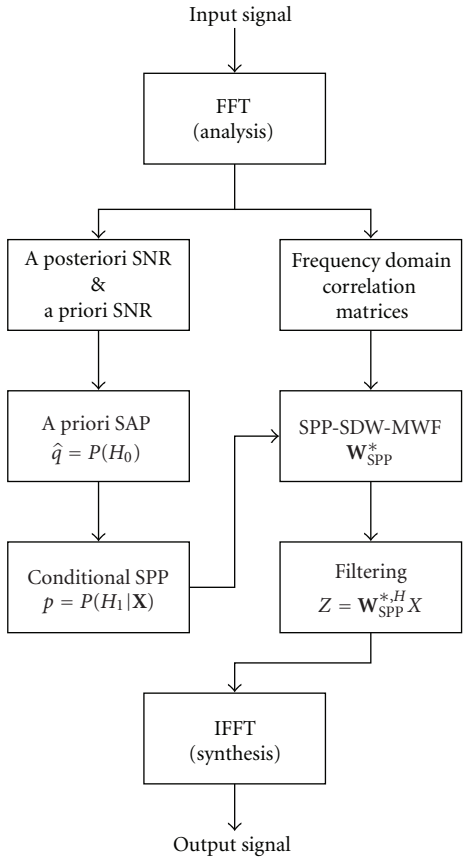

Figure 7: Block diagram of the proposed SDW-MWF$_{\text{SPP}}$ incorporating the conditional SPP.

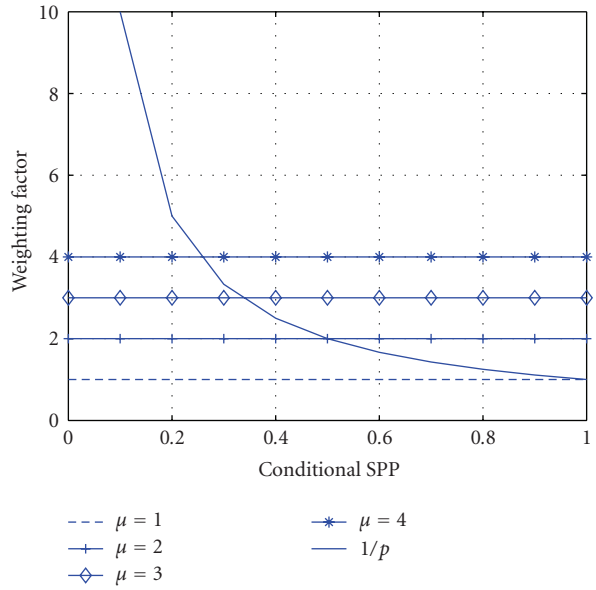

Figure 8: Speech presence probability-based weighting factor compared to a fixed weighting factor.

noise reduction when p is small, that is, for noise dominant segments, and less noise reduction when p is high, that is, for speech dominant segments, as shown in Figure 8 (solid line). This concept is compared to a fixed weighting factor μ used in a traditional SDW-MWF$_\mu$ that does not take speech presence or absence into account as follows.

(i) If $p = 0$, that is, when the probability that speech is present is zero, the SDW-MWF$_{SPP}$ attenuates the noise by applying $\mathbf{W}^* \leftarrow 0$.

(ii) If $p = 1$, that is, when the probability that speech is presence is one, the SDW-MWF$_{SPP}$ solution corresponds to the MWF solution ($\mu = 1$).

(iii) If $0 < p < 1$, there is a trade-off between noise reduction and speech distortion based on the conditional SPP.

5.1. Undesired Noise Modelling. The problem with SDW-MWF$_{SPP}$ derived in (23) is that the inverse of the conditional SPP is used, which can cause large fluctuations in different frequency bands especially if the weighting factor $1/p$ is used, as shown in Figure 7. For example, if the conditional SPP shown in Figure 5 is used, the SDW-MWF$_{\text{SPP}}$ will apply a NR corresponding to $\mu = 1$ below 2000 Hz, and between 2000 Hz, and 4500 Hz the NR will be much larger. This transition between low and high NR in different frequency bands can cause speech distortion or musical noise.

It is also worth noting that in the derivation of the SDW-MWF$_{\text{SPP}}$ the term $(1 - p) = P(H_0 \mid X_i)$ is not present in (22) anymore. This can be explained by the fact that the SDW-MWF estimates the speech component in one of the microphones under hypothesis H_1 while under hypothesis H_0 the noise reduction filter is set to zero. In [18] the gain function is similarly derived under hypothesis H_1, which is due to the fact that the method aims to provide an estimate of the clean speech spectrum, so that when the speech is absent the gain is set to zero. This property negatively affects the processing of the noise-only bins which results in undesired modelling of the noise making the residual noise sounds unnatural.

5.2. Combined Solution. In [12, 17] a lower threshold is introduced for the gain under hypothesis H_0. This lower threshold is based on subjective criteria for the noise naturalness. Applying a constant attenuation when the speech is absent results in a uniform noise level, and therefore any undesired noise modelling can be avoided so that the naturalness of the residual noise can be retained.

Following the concept with the lower threshold a solution is proposed that in one extreme case corresponds to the SDW-MWF$_{\text{SPP}}$ and in the other extreme case corresponds to a traditional SDW-MWF$_\mu$. The combined solution can then be written as

$$\mathbf{W}^*_{\text{SPP}} = \left(\mathbf{R}^s + \left(\frac{1}{\alpha(1/\mu) + (1-\alpha)p}\right)\mathbf{R}^n\right)^{-1} \mathbf{R}^s \mathbf{e}_1, \quad (24)$$

where μ in this case is the constant attenuation factor, and α is a trade-off factor between SDW-MWF$_\mu$ and SDW-MWF$_{\text{SPP}}$.

The weighting factor for the combined solution is shown in Figure 9 for $\alpha = 0.5$ and for different values of μ. The concept then goes as follows.

(i) If $\alpha = 1$, the solution corresponds to a traditional SDW-MWF$_\mu$ given in (6).

(ii) If $\alpha = 0$, the solution corresponds to the SDW-MWF$_{\text{SPP}}$ given in (23).

(iii) If $0 < \alpha < 1$, there is a trade-off between the two solutions based on μ and α and p given in (24).

(iv) If $p = 0$, that is, when the probability that speech is present is zero, the SDW-MWF$_{\text{SPP}}$ attenuates the noise by applying a constant weighting, that is, μ/α corresponding to the desired lower threshold.

The conditional SPPs for $\zeta_{\min} = 0.1$ and $\zeta_{\max} = 0.3162$ in Figures 5 and 6 for the combined solution are shown in Figures 10 and 11. When α is increased, the solution gets closer to the standard SDW-MWF$_\mu$ ($\mu = 2$). The importance of SDW-MWF$_{\text{SPP}}$ is that different amount of NR can be applied to the speech dominant segments and to the noise dominant segments. With the combined solution the overall amount of NR might not exceed SDW-MWF$_\mu$, but the distinction between speech and noise is the important part in order to enhance speech dominant segments and further suppress the noise dominant segments. Increasing α limits the distortion but in the same time also limits the NR in a narrow frequency band; that is, the ratio between the speech dominant segments and the noise dominant segments are reduced. Furthermore the weak high-frequency component might also be less emphasized since less NR is applied to frequencies prior to the weak high-frequency component; see Figures 10 and 11. This combined solution does not only offer a flexibility between the SDW-MWF$_{\text{SPP}}$ and a traditional SDW-MWF$_\mu$. In this case α effectively determines the dynamics of the SDW-MWF$_{\text{SPP}}$ and the degree of nonlinearity in the weighting factor.

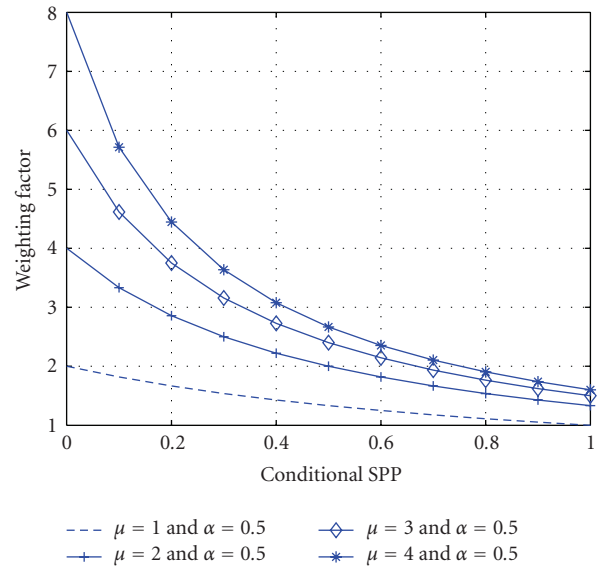

Figure 9: Speech presence probability-based weighting factor with a lower threshold.

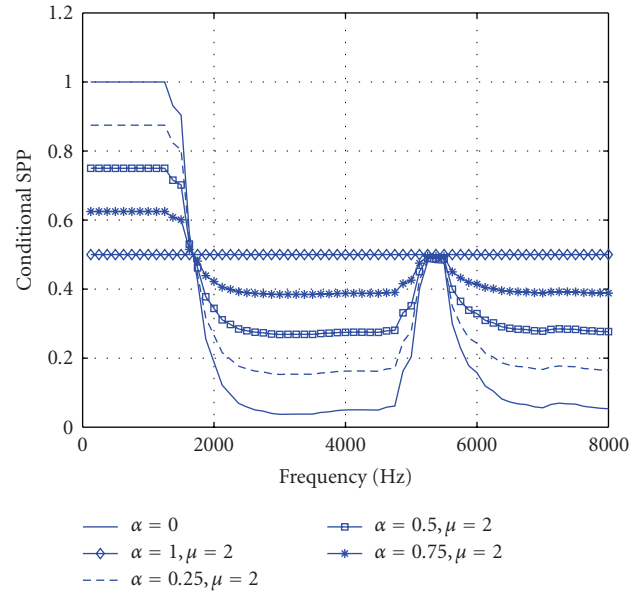

Figure 10: Conditional SPP for the combined solution with high-speech presence at low frequency.

6. Experimental Results

In this section, experimental results for the proposed SDW-MWF$_{\text{SPP}}$ ($\alpha = 0$) are presented and compared to a traditional SDW-MWF$_\mu$ ($\alpha = 1$). In-between solutions of these two approaches are also presented.

6.1. Experimental Set-up. 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 $T_{60} = 0.21$ seconds. The speech is located at 0°, and the two multitalker babble noise sources are located at 120° and 180°. The speech signals consist of male sentences from the HINT-database [19], and the noise signals consist of a multi-talker babble from Auditec [20]. The speech signals are sampled at 16 kHz and are concatenated as shown in Figure 2. For the estimation of the second-order statistics access to a perfect VAD was assumed. An FFT length of 128 with 50% overlap was used. Table 1 shows the parameters used in the estimation of the conditional SPP.

6.2. Performance Measures. To assess the noise reduction performance the intelligibility-weighted signal-to-noise ratio (SNR) [21] is used which is defined as

$$\Delta\text{SNR}_{\text{intellig}} = \sum_i I_i \left(\text{SNR}_{i,\text{out}} - \text{SNR}_{i,\text{in}}\right), \tag{25}$$

where I_i is the band importance function defined in [22], and where SNR$_{i,\text{out}}$ and SNR$_{i,\text{in}}$ represent the output SNR and the input SNR (in dB) of the ith band, respectively. For measuring the signal distortion a frequency-weighted

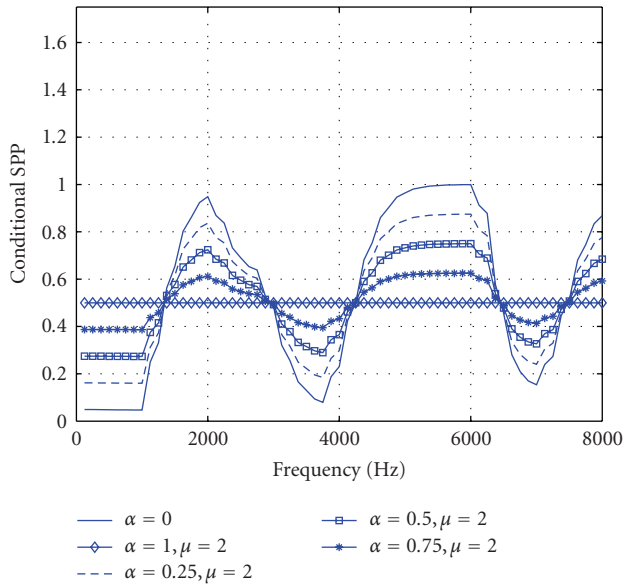

Figure 11: Conditional SPP for the combined solution with two distinct speech dominant segments.

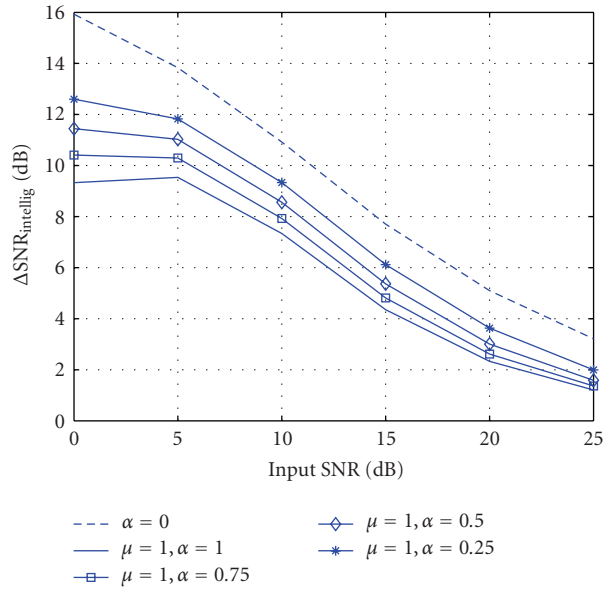

Figure 12: SNR improvement for SDW-MWF$_{\text{SPP}}$ ($\alpha = 0$) and SDW-MWF$_{\mu}$ ($\alpha = 1$) with $\mu = 1$ at different input SNRs.

Table 1: Parameters used for the estimation of the conditional SPP.

$\beta = 0.7$	$\rho = 0.95$	$\kappa = 0.98$
$\omega_{\text{local}} = 1$	$\zeta_{\min} = -10\,\text{dB}\ (0.1)$	$\zeta_{p\ \min} = 4\,\text{dB}$
$\omega_{\text{global}} = 10$	$\zeta_{\max} = -5\,\text{dB}\ (0.3162)$	$\zeta_{p\ \max} = 10\,\text{dB}$

log-spectral signal distortion (SD) is used defined as

$$\text{SD} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\int_{f_l}^{f_u} w_{\text{ERB}}(f)\left(10\log_{10}\frac{P^s_{\text{out},k}(f)}{P^s_{\text{in},k}(f)}\right)^2 df}, \quad (26)$$

where K is the number of frames, $P^s_{\text{out},k}(f)$ is the output power spectrum of the kth frame, $P^s_{\text{in},k}(f)$ is the input power spectrum of the kth frame, and f is the frequency index. The SD measure is calculated with a frequency-weighting factor $w_{\text{ERB}}(f)$ giving equal weight for each auditory critical band, as defined by the equivalent rectangular bandwidth (ERB) of the auditory filter [23].

6.3. SDW-MWF$_{\text{SPP}}$ versus SDW-MWF$_{\mu}$. The performance of the SDW-MWF$_{\text{SPP}}$ ($\alpha = 0$) and SDW-MWF$_{\mu}$ ($\mu = 1$ and 2) is evaluated for different input SNRs ranging from 0 dB to 25 dB. The combined solution is evaluated for different values of $\alpha = 0.25, 0.50$, and 0.75, since this provides a trade-off between a traditional SDW-MWF$_{\mu}$ and the proposed SDW-MWF$_{\text{SPP}}$.

The SNR improvement and SD for different input SNRs are shown in Figures 12–15. It is clear that when $\alpha \to 0$ the SNR improvement is larger, but at the same time the SD also increases. When α is increased, the SNR improvement decreases, and at the same time the SD also decreases. It was found that an α value around 0.25 to 0.5 reduces the signal distortion significantly, but this obviously comes at the cost of less improvement in SNR. As mentioned

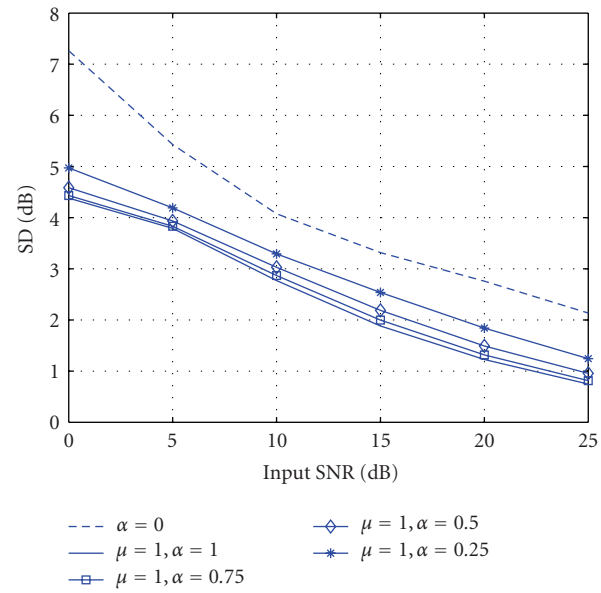

Figure 13: Signal distortion for SDW-MWF$_{\text{SPP}}$ ($\alpha = 0$) and SDW-MWF$_{\mu}$ ($\alpha = 1$) with $\mu = 1$ at different input SNRs.

the goal of SDW-MWF$_{\text{SPP}}$ is not to necessarily outperform SDW-MWF$_{\mu}$ in terms of SNR or SD. The motivation behind SDW-MWF$_{\text{SPP}}$ is to apply less NR to speech dominant segments and more NR to noise dominant segments. Therefore the overall weighting factor in the combined solution might not be higher than SDW-MWF$_{\mu}$. Actually when $\mu = 2$ and $\alpha = 0.5$, the NR applied when the conditional SPP is larger than 0.5 is lower than $\mu = 2$; see Figure 9 (solid line).

6.4. Residual Noise. A reason for the increased SD can be caused by the sharp transition between speech dominant

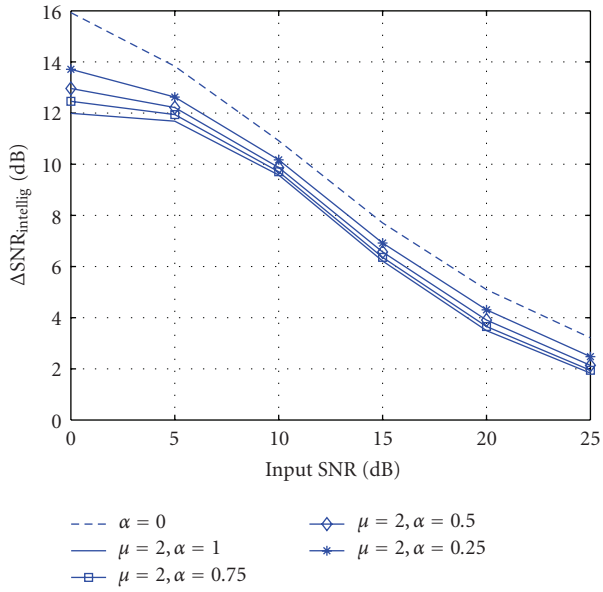

FIGURE 14: SNR improvement for SDW-MWF$_{SPP}$ ($\alpha = 0$) and SDW-MWF$_{\mu}$ ($\alpha = 1$) with $\mu = 2$ at different input SNRs.

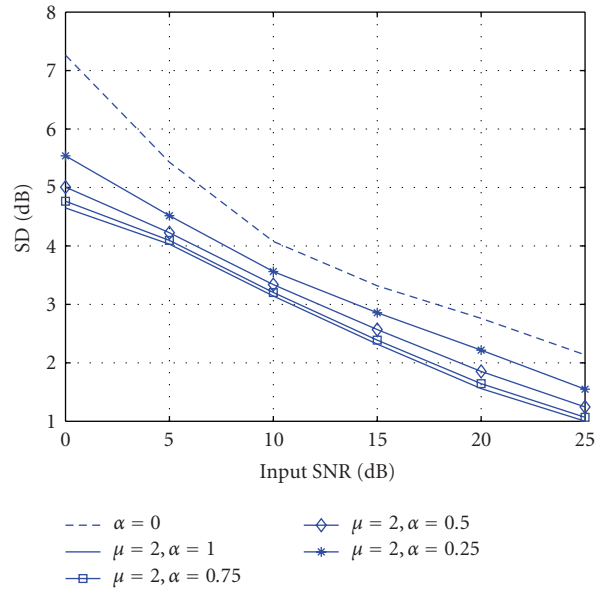

FIGURE 15: Signal distortion for SDW-MWF$_{SPP}$ ($\alpha = 0$) and SDW-MWF$_{\mu}$ ($\alpha = 1$) with $\mu = 2$ at different input SNRs.

segments and noise dominant segments; see Figures 5 and 6 and Section 5.1. A softer transition in this case will probably be desired, for example, by applying smoothing to the conditional SPP or by modifying the threshold functions in (16) and (19).

One way of interpreting the results from the SD measure is to look at the residual noise. When $\alpha \to 0$, the musical noise phenomenon occurs, while it is less significant when $\alpha \to 1$ which partly can be supported by the SD measure shown in Figure 13. Using an α value around 0.25 to 0.5 reduces the musical noise and makes the noise sound more natural. It is also observed that the noise modelling of the residual noise is more significant in the noise-only periods where the update of the SDW-MWF$_{SPP}$ occurs, see Figure 2. The goal of SDW-MWF$_{SPP}$ is to attenuate the noise dominant segments more compared to speech dominant segments. The question is still whether this SD measure has any effect on the speech intelligibility. This may not be the case if only the noise dominant segments are attenuated more compared to the speech dominant segments. If the conditional SPP is accurate, the speech dominant segments can be made more significant compared to the noise dominant segments, especially if the NR is able to reduce the noise in a narrow frequency band. The benefit of this concept is still something that needs to be analyzed.

Musical noise is not an effect normally encountered in multi-channel noise reduction. This typically appears in single-channel noise reduction that is based on short-time spectral attenuation. Increasing α reduces the musical noise, which basically means that the fast tracking of speech presence in each frequency bin and each frame is constrained. The function of α is to a trade-off between a traditional SDW-MWF$_{\mu}$, that is, a linear slow time-varying system and a SDW-MWF$_{SPP}$, that is, a nonlinear fast time-varying system.

7. Conclusion

In this paper an SDW-MWF$_{SPP}$ procedure has been presented that incorporates the conditional SPP. A traditional SDW-MWF$_{\mu}$ uses a fixed parameter to a trade-off between noise reduction and speech distortion without taking speech presence into account. Incorporating the conditional SPP in SDW-MWF allows to exploit the fact that speech may not be present at all frequencies and at all times, while the noise can indeed be continuously present. This concept allows the noise to be reduced in a narrow frequency band based on the conditional SPP. 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. A combined solution is also proposed that in one extreme case corresponds to an SDW-MWF$_{SPP}$ and in the other extreme case corresponds to a traditional SDW-MWF$_{\mu}$ solution. In-between solutions correspond to a trade-off between the two extreme cases.

The SDW-MWF$_{SPP}$ is found to significantly improve the SNR compared to a traditional SDW-MWF$_{\mu}$. The SNR improvement however comes at the cost of audible musical noise, and here the in-between solutions offer a way to reduce the musical noise while still maintaining an SNR improvement that is larger than SDW-MWF$_{\mu}$. The explanation of this is due to the fact that a traditional SDW-MWF$_{\mu}$ implementation is a linear filter and is based on a long-term average of the spectral and spatial signal characteristics, whereas the SDW-MWF$_{SPP}$ has a weighting factor changing on a faster dynamic for each frequency bin and each frame, which corresponds better to the nonstationarity of the speech and the noise characteristics.

Acknowledgments

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 contact 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 the authors.

References

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

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using optimal non-linear spectral amplitude estimation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '83)*, vol. 8, pp. 1118–1121, April 1983.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[6] O. L. I. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

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

[8] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[9] 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. 49, no. 7-8, pp. 636–656, 2007.

[10] 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, 2005.

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

[12] 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, 2002.

[13] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[14] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. Jensen, "Variable speech distortion weighted multichannel wiener filter based on soft output voice activity detection for noise reduction in hearing aids," in *Proceedings of the 11th International Workshop on Acoustic Echo and Noise Control (IWAENC '08)*, Seattle, Wash, USA, 2008.

[15] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, Fla, USA, 2007.

[16] H. Levitt, "Noise reduction in hearing aids: a review," *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, pp. 111–121, 2001.

[17] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[18] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, vol. 2, pp. 789–792, Phoenix, Ariz, USA, March 1999.

[19] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, 1994.

[20] Auditec, "Auditory Tests (Revised), Compact Disc, Auditec, St. Louis," St. Louis, 1997.

[21] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, 1993.

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

[23] B. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, New York, NY, USA, 5th edition, 2003.