COMPARISON OF FREQUENCY DOMAIN NOISE REDUCTION STRATEGIES BASED ON MULTICHANNEL WIENER FILTERING AND SPATIAL PREDICTION

Bram Cornelis, Marc Moonen*

Katholieke Universiteit Leuven
Dept. of Electrical Engineering (ESAT–SCD)
Kasteelpark Arenberg 10,
3001 Leuven, Belgium

Jan Wouters

Katholieke Universiteit Leuven
Dept. of Neurosciences, ExpORL
Herestraat 49/721,
3000 Leuven, Belgium

ABSTRACT

In this paper two multichannel noise reduction strategies are compared in the context of binaural hearing aids. Recently a novel noise reduction method based on spatial-temporal prediction (STP) was introduced which showed an improvement over methods based on multichannel Wiener filtering, although at the cost of a higher computational complexity. Whereas this new method operates in the time domain, hearing aids typically demand faster frequency domain implementations. In this paper we therefore propose a frequency domain equivalent of the STP method. The performance of the new so-called spatial prediction (SP) method will be compared to a frequency domain implementation of the speech distortion weighted multichannel Wiener filter (SDW-MWF), theoretically as well as based on simulations with a binaural hearing aid configuration. It will be shown that the frequency domain SP method still achieves some improvement over the SDW-MWF, at the cost of higher computational complexity.

Index Terms— binaural hearing aids, noise reduction, multichannel Wiener filtering, spatial-temporal prediction

1. INTRODUCTION

For several years now noise reduction has been an active area of research with applications such as speech communications and especially digital hearing aids. The first procedures applied in hearing aids made use of a single microphone and assumed additive noise [1]. In recent years however, hearing aids have been fitted with multiple microphones so that multi-microphone beamforming techniques can be used. These have the potential to achieve a higher SNR improvement than in the single microphone case, while still keeping the speech distortion at an acceptable level.

A popular class of multi-microphone noise reduction procedures is based on the generalized sidelobe canceller (GSC) structure [2]. The initial procedure assumed free-field propagation, but this was extended to arbitrary transfer functions in the transfer function linearly constrained minimum variance (TF-LCMV) method [3]. A more recent method referred to as speech distortion weighted multichannel Wiener filtering (SDW-MWF) [4] makes use of the speech and noise correlation matrices to obtain filters that minimize the mean square error (MSE) of the residual noise energy, where a trade-off parameter controls the speech distortion. The SDW-MWF eliminates the need for a fixed beamformer preprocesser (as in the GSC structure), hence offers a very promising alternative to the GSC.

In [5, 6] a novel multi-microphone noise reduction procedure is proposed which also makes use of the speech and noise correlation matrices. The signals are preprocessed by a spatial-temporal prediction matrix and we will therefore refer to the method as the STP method. By virtue of this preprocessing minimum speech distortion can be imposed. In [6] this procedure was also compared to the multichannel Wiener filter. It was reported that the STP method is more robust against errors in the estimation of the speech correlation matrix, so that a significant improvement over the MWF can be obtained.

The STP method shows this improvement especially when a large microphone array is used. For a hearing aid application however, typically only 2 or 3 microphones are used. This number could be doubled in the future when binaural hearing aids (bilateral hearing aids connected by a wireless link) can exchange microphone signals, be it that the array size would still be only moderate. Another characteristic of hearing aids is the limited memory and processing power. Whereas the method in [5] operates in the time domain, hearing aids typically require a faster frequency domain implementation. In this view, the basic question arises whether the new method would be better suited than the SDW-MWF for a binaural hearing aid.

In this paper, the principle introduced in [5] will be applied in a frequency domain method and tested on a binaural hearing aid configuration. This frequency domain method only uses the spatial correlations of the speech signals and will therefore be referred to as the Spatial Prediction (SP) method. Although a smaller SNR improvement will be obtained by dropping the temporal correlations out of the prediction, the SP method becomes computationally feasible for an implementation on hearing aids.

The theoretical analysis will show that the SP method is highly related to the TF-LCMV method and is also connected to the SDW-MWF method. The optimal filters and output SNR of both procedures will be derived under the assumption of a single speech source. We finally conclude with simulations conducted on a binaural hearing aid configuration, and we will show that the SP method can achieve a performance improvement over a frequency domain implementation of the SDW-MWF even for a binaural hearing aid configuration, be it at the cost of a higher computational complexity.

*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-AMBioRICS 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.

2. CONFIGURATION AND NOTATION

2.1. Microphone signals and output signals

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. Assuming that the speech and the noise components are uncorrelated, $\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_v$. The noise reduction algorithms considered here are based on a linear filtering of the microphone signals. The microphone signals are filtered by a filter $\mathbf{W}$ so that an output signal Z is obtained as $Z = \mathbf{W}^H\mathbf{Y}$.

2.2. Single speech source assumption

In the case of a single speech source, the speech signal vector can be modelled as

$$\mathbf{X} = \mathbf{A}S\,, \tag{3}$$

where $\mathbf{A}$ 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-1 matrix, i.e.

$$\mathbf{R}_x = P_s\mathbf{A}\mathbf{A}^H\,, \tag{4}$$

with $P_s = \mathcal{E}\{|S|^2\}$ the power of the speech signal. The single speech source assumption will be used in the theoretical analysis.

3. NOISE REDUCTION: SDW-MWF AND SPATIAL PREDICTION (SP)

In this section, we will compare two frequency domain noise reduction procedures. We will define one of the microphones as the reference microphone, and define the error signal as the difference between the output signal and the (unknown) speech component of the reference microphone, i.e.

$$\begin{aligned} E &= Z - X_{\text{ref}}\,, &(5)\\ &= (\mathbf{W}-\mathbf{u})^H\mathbf{X} + \mathbf{W}^H\mathbf{V}\,, &(6)\\ &= E_x + E_v\,, &(7)\end{aligned}$$

where $\mathbf{u}$ is a vector with one entry equal to one and all other entries equal to zero, so that $\mathbf{u}^H\mathbf{X} = X_{\text{ref}}$. The error signal is split into two components, namely the speech distortion error E_x and the residual noise error E_v. It is possible to define MSE cost functions for the filter $\mathbf{W}$ based on these error signals:

$$J_v(\mathbf{W}) = \mathcal{E}\{E_vE_v^*\} = \mathbf{W}^H\mathbf{R}_v\mathbf{W}\,, \tag{8}$$

and

$$J_x(\mathbf{W}) = \mathcal{E}\{E_xE_x^*\} = (\mathbf{W}-\mathbf{u})^H\mathbf{R}_x(\mathbf{W}-\mathbf{u})\,. \tag{9}$$

3.1. Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF)

Noise reduction can be obtained by minimizing the residual noise error cost function (8) with respect to $\mathbf{W}$. However, the speech distortion error (9) can then become arbitrarily large. Therefore, a constraint will be imposed to keep the speech distortion error under a threshold T. This leads to the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \mathbf{W}^H\mathbf{R}_v\mathbf{W} &(10)\\ \text{s.t.} \quad & (\mathbf{W}-\mathbf{u})^H\mathbf{R}_x(\mathbf{W}-\mathbf{u}) \le T\,. &(11)\end{aligned}$$

By introducing the Lagrange multiplier λ, an equivalent unconstrained problem is obtained:

$$\min_{\mathbf{W}} \quad \mathbf{W}^H\mathbf{R}_v\mathbf{W} + \lambda\left[(\mathbf{W}-\mathbf{u})^H\mathbf{R}_x(\mathbf{W}-\mathbf{u}) - T\right]\,. \tag{12}$$

It can then be shown that the optimal filter is equal to:

$$\boxed{\mathbf{W}_{SDW} = \left(\mathbf{R}_x + \mu\mathbf{R}_v\right)^{-1}\mathbf{R}_x\mathbf{u}} \tag{13}$$

where $\mu = \frac{1}{\lambda}$. This is referred to as the speech-distortion weighted multichannel Wiener filter (SDW-MWF) [4]. The parameter μ allows a trade-off between speech distortion and noise reduction.

3.2. Spatial Prediction

In [5], a novel time domain noise reduction procedure was proposed. By making use of the temporal and spatial correlations of the speech signals, spatial-temporal prediction matrices were estimated, which allowed to impose the speech distortion error to be zero. Here this principle will be adopted in a frequency domain approach.

The N speech components can be related to the reference speech signal: $X_n = \mathcal{H}_{n,ref}\,X_{\text{ref}}$, for $n = 1...N$, so that

$$\mathbf{X} = \begin{bmatrix} \mathcal{H}_{1,ref} \\ \vdots \\ \mathcal{H}_{N,ref} \end{bmatrix} X_{\text{ref}} = \mathbf{H}\,X_{\text{ref}}. \tag{14}$$

In contrast to the time domain approach in [5], we now only make use of the spatial correlations between the speech components, hence only a spatial prediction is performed. The spatial prediction vector $\mathbf{H}$ can be found in the Wiener sense, i.e. by minimizing

$$\min_{\mathbf{H}} \; \mathcal{E}\left\{(\mathbf{X} - \mathbf{H}\,X_{\text{ref}})^H\,(\mathbf{X} - \mathbf{H}\,X_{\text{ref}})\right\} \tag{15}$$

which leads to

$$\boxed{\mathbf{H} = \frac{1}{\mathbf{u}^H\mathbf{R}_x\mathbf{u}}\,\mathbf{R}_x\mathbf{u}} \tag{16}$$

which means one column of the speech correlation matrix is selected and divided by the speech component power in the reference microphone. Using (14), the speech distortion error E_x can be written as

$$E_x = (\mathbf{W}-\mathbf{u})^H\mathbf{X} = (\mathbf{W}^H\mathbf{H} - 1)X_{\text{ref}} \tag{17}$$

so that $E_x = 0$ if $\mathbf{W}^H\mathbf{H} = 1$.

In contrast to the optimization problem of the SDW-MWF (11), where the speech distortion was controlled by a soft constraint, we

can now impose the speech distortion to be zero, which leads to the following expression:

$$\min_{\mathbf{W}} \quad \mathbf{W}^H \mathbf{R}_v \mathbf{W} \tag{18}$$

$$\text{s.t.} \quad \mathbf{W}^H \mathbf{H} = 1 \; . \tag{19}$$

It is easily shown that the optimal filter is equal to

$$\boxed{\mathbf{W}_{SP} = \frac{1}{\mathbf{H}^H \mathbf{R}_v^{-1} \mathbf{H}} \mathbf{R}_v^{-1} \mathbf{H}} \; . \tag{20}$$

3.3. Special case: single speech source

We can now use the single source assumption to analyze and compare the theoretical performance of both procedures. By plugging (3) and (4) into (13), it can be shown that the optimal filter for the SDW-MWF is equal to

$$\mathbf{W}_{SDW} = \frac{P_s}{\mu + \rho} \mathbf{R}_v^{-1} \mathbf{A} A_{ref}^* \tag{21}$$

with $A_{ref}^* = \mathbf{A}^H \mathbf{u}$ and $\rho = P_s \mathbf{A}^H \mathbf{R}_v^{-1} \mathbf{A}$.

The spatial prediction vector (16) becomes equal to

$$\mathbf{H} = \frac{1}{A_{ref}} \mathbf{A} \tag{22}$$

in the single speech source case. Remarkably, this is the transfer function ratio that is also estimated in the TF-LCMV method [3], where the transfer function ratio is used to calibrate the fixed beamformer in the preprocessing step of the GSC structure. The implementation of the SP method described here does not use the GSC structure however, but calculates the filters based on the estimated speech and noise correlation matrices, as in the implementation of the SDW-MWF.

By plugging (3), (4) and (22) into (20), it can be shown that the optimal filter for the SP method is equal to:

$$\mathbf{W}_{SP} = \frac{P_s}{\rho} \mathbf{R}_v^{-1} \mathbf{A} A_{ref}^* \tag{23}$$

which is also the filter which is obtained by the TF-LCMV, after convergence.

Remarkably, the optimal filters of SDW-MWF (21) and SP (23) are parallel and can be related by a scalar factor as

$$\boxed{\mathbf{W}_{SDW} = \mathbf{W}_{SP} \, \frac{\rho}{\mu + \rho} \; .} \tag{24}$$

As a consequence, the SDW-MWF and SP obtain the same output SNR (per frequency bin), i.e.

$$SNR_{out,f} = \frac{\mathbf{W}^H \mathbf{R}_x \mathbf{W}}{\mathbf{W}^H \mathbf{R}_v \mathbf{W}} = P_s \; \mathbf{A}^H \mathbf{R}_v^{-1} \mathbf{A} = \rho \; . \tag{25}$$

Note that this is also independent of the trade-off parameter μ. The μ parameter will have an effect on the obtained broadband SNR's however, where a higher value of μ leads to more noise reduction.

4. SIMULATIONS

4.1. Setup

An interesting application for a noise reduction procedure is a binaural hearing aid configuration, i.e. two hearing aids connected by a wireless link. In this paper we will assume that the link is ideal in terms of bandwidth and power consumption. We therefore assume that all microphone signals are available as input to the noise reduction procedure, where 2 microphones are at the left ear and 2 at the right ear, giving a total of $N = 4$. The binaural procedure produces a stereo output, but in these tests only the output for the left ear will be shown. The left-front microphone is chosen as the reference microphone.

Head-related transfer functions (HRTF's) were measured in a reverberant room (reverberation time RT_{60} = 500 ms) on a dummy-head, so that the head-shadow effect is taken into account. To generate the microphone signals, the noise and speech signals are convolved with the HRTF's corresponding to their angles of arrival, before being added together. 11 different speech-noise configurations were tested. The azimuthal angles (defined clockwise with 0° as frontal direction) of the speech and noise sources are varied. The speech signal is at 0° in all scenario's except the last. In the first 6 scenario's there is a single noise source at $\theta_v = 60^\circ, 90^\circ, 120^\circ, 180^\circ, 270^\circ, 300^\circ$. Scenario 7, 8 and 9 feature 2 noise sources at $\theta_v = [-60^\circ 60^\circ]$, $\theta_v = [-120^\circ 120^\circ]$ and $\theta_v = [120^\circ 210^\circ]$. Scenario 10 features 4 noise sources at $\theta_v = [60^\circ 120^\circ 180^\circ 210^\circ]$. In scenario 11 the speech source is at 90° and the noise source at 180°. For the noise signal(s) multitalker babble noise is used, the speech signal consists of 4 sentences of the Hearing In Noise Test (HINT) list. The microphone signals have a total length of 26 s and are sampled at 20480 Hz.

A batch procedure was implemented where the speech and noise correlation matrices are estimated off-line using the 26s microphone signals. The speech correlation matrix is estimated as $\mathbf{R}_x = \mathbf{R}_y - \mathbf{R}_v$. In practice, a voice activity detector (VAD) has to be implemented to distinguish between moments where speech and noise are both active ($\mathbf{R}_y$ is updated) and moments where only noise is active ($\mathbf{R}_v$ is updated), but here a perfect VAD is assumed.

The optimal SDW-MWF and SP filters are then found by plugging the correlation matrix estimates into (13) and (20).

4.2. Single Frequency SNR

In a first experiment, the output SNR per frequency bin (25) is calculated in the realistic setup described in the previous section, for SDW-MWF ($\mu = 1$ and $\mu = 5$) and SP. Only the result for scenario 10 (1 speech source, 4 noise sources) is shown here. The SNR is calculated according to the left expression in (25) for every frequency bin (for L=128 frequency bins). The results are shown in figure 1.

Although the theoretical analysis showed that the output SNR is independent of μ, the SDW-MWF with $\mu = 5$ obtains a higher SNR than the SDW-MWF with $\mu = 1$. The SP procedure achieves a performance comparable to the SDW-MWF with $\mu = 5$. The reason for this difference can be explained by studying the SDW-MWF and SP optimal filters (13) and (20). The speech and noise correlation matrices are used in different ways in these formula's. As the noise signal is typically more stationary than the speech signal, the estimation of the noise correlation matrix is easier than the estimation of the speech correlation matrix. In addition, the speech correlation matrix can only be estimated indirectly ($\mathbf{R}_x = \mathbf{R}_y - \mathbf{R}_v$), which further increases the estimation error. Because of this estimation error, the performance will degrade.

In the SDW-MWF formula (13), the entire speech correlation

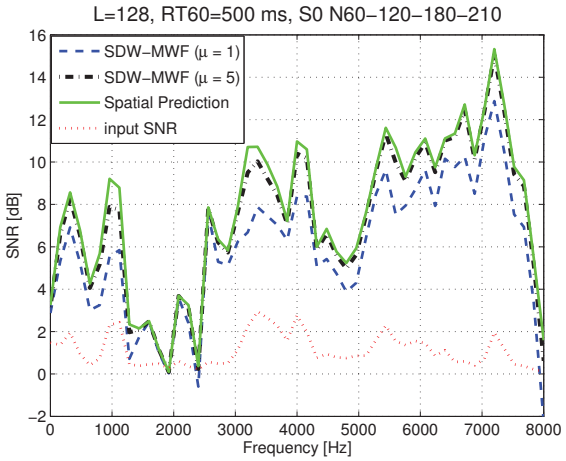


Fig. 1. *Output SNR per frequency bin for SDW-MWF and SP*

matrix is added to a scaled version of the noise correlation matrix (with factor μ) and then inverted. For higher values of μ, the estimation error of the speech correlation matrix will have a smaller effect on the performance because the term with the noise correlation matrix is dominant. This can explain why $\mu = 5$ has a higher SNR than the case $\mu = 1$.

The SP procedure does not use the entire speech correlation matrix in (20), but only a column in the spatial prediction step (16). This explains why the performance is better than in the SDW-MWF case.

4.3. SI weighted broadband SNR and SD

In figure 2, the broadband performances of the SDW-MWF and SP procedures for the 11 speech-noise configurations are shown. To assess the improvement in speech intelligibility (SI), the SI-weighted SNR improvement and SI-weighted speech distortion (SD) defined in [7] are used.

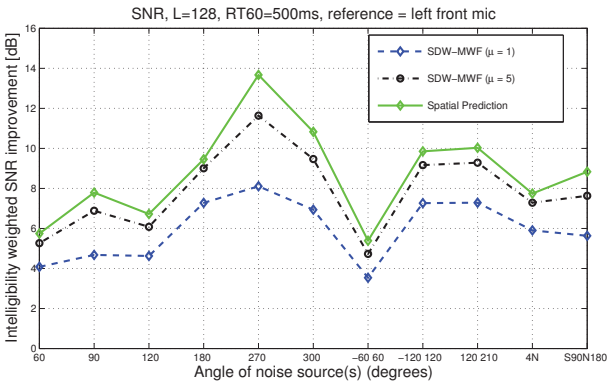
The SDW-MWF with $\mu = 1$ has a low speech distortion, but also obtains the smallest SNR improvement. With $\mu = 5$, we can put more emphasis on noise reduction and this indeed results in a larger SNR improvement. However, the speech distortion then becomes very large. The SP procedure has a small SD which is comparable to the SDW-MWF ($\mu = 1$), and still achieves a large SNR improvement which is slightly better than the SDW-MWF with $\mu = 5$.

5. CONCLUSION

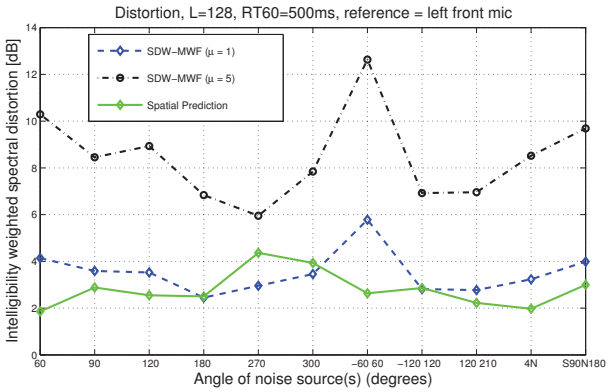
In this paper, we have shown through experiments that the SP procedure enables a performance improvement both in SNR and in SD. The SP formula's (16) and (20) are more robust to estimation errors in the speech correlation matrix, which explains this improvement. On the other hand, the SP procedure is computationally more complex than the SDW-MWF, so that it will depend on the available processing power whether SP is preferred over SDW-MWF. Extensive perceptual tests should then also be performed to see if the SNR and SD improvements indeed correspond to an improvement in speech intelligibility.

6. REFERENCES

[1] C.P. Loizou, *Speech enhancement: Theory and Practice*, CRC press, New York, USA, 2007.



(a) SNR improvement



(b) Speech Distortion (SD)

Fig. 2. *SI-weighted SNR improvement (a) and SD (b) for 11 speech-noise configurations, for SDW-MWF and SP.*

[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] S. Gannot, D. Burshtein, and E. Weinstein, "Signal Enhancement Using Beamforming and Non-Stationarity with Applications to Speech," *IEEE Trans. Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[4] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, *Speech Distortion Weighted Multichannel Wiener Filtering Techniques for Noise Reduction*, chapter 9 in "Speech Enhancement" (J. Benesty, J. Chen, S. Makino, Eds.), pp. 199–228, Springer-Verlag, 2005.

[5] Jingdong Chen, J. Benesty, and Yiteng Huang, "A minimum distortion noise reduction algorithm with multiple microphones," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 3, pp. 481–493, March 2008.

[6] Yiteng Huang, J. Benesty, and Jingdong Chen, "Analysis and comparison of multichannel noise reduction methods in a common framework," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 957–968, July 2008.

[7] 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 and Audio Processing*, vol. 13, no. 4, pp. 487–503, July 2005.