Perceptually-inspired Processing for Multichannel Wiener Filter

Jorge I. Marin-Hurtado1,2, David V. Anderson1

1School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA
2Department of Electronics Engineering, Universidad del Quindío, Armenia, Q., Colombia

{jmih3, anderson}@gatech.edu

INTERSPEECH 2011

Abstract

Binaural noise-reduction techniques based on Multichannel Wiener filter (MWF) have been reported as promissory candidates to be used in binaural hearing aids because of their effective SNR improvement at any arbitrary direction of arrival of the target signal and the preservation of localization cues. There are different MWF techniques derived in the FFT domain. The use of an FFT-based processing involves two important challenges for the real-time implementation of these techniques in a digital hearing: high computational cost and processing delay. To reduce computational cost and processing delay without degrading the SNR improvement and sound quality, this paper proposes the use of an auditory representation instead of an FFT representation. The proposed processing shows significant advantages over an FFT-based processing: reduction of the computational cost and processing delay, and improvement of the output SNR and sound quality.

Index Terms: Noise Reduction, Multichannel Wiener Filter, Auditory Filter-bank, Binaural Hearing Aids

1. Introduction

Binaural hearing aids use a wireless link to exchange the signals received at each side, allowing the implementation of sophisticated binaural noise-reduction strategies [1, 2, 3]. In the past decade, researchers have shown the benefits of binaural processing over monaural processing to deal with very hostile environments such as babble noise, to provide enhancement of target signals coming from any arbitrary direction of arrival, and to preserve the localization cues. A recent study compared different binaural noise-reduction techniques based on scene analysis, spectral subtraction, adaptive beamforming, and multichannel Wiener filter (MWF), and concluded that the MWF approaches offer the best signal-to-ratio (SNR) improvement and output sound quality [4].

Wiener filter coefficients are estimated by minimizing a cost function, typically the mean square error between the desired signal (clean speech) and the Wiener filter output. The multisensor variant of a Wiener filtering technique is called multichannel Wiener filter (MWF). In the MWF framework, the signals of all channels are analyzed in the FFT domain. For each frequency bin, the signal and noise statistics are estimated and used to compute the weights that filter out the noise.

Although MWF methods have been identified as a good choice to be used in binaural hearing aids, a practical real-time implementation of these methods involves some challenges. First, the implementation of FFT-based MWF involves high computational resources, since the signal and noise correlation matrices as well as the weights need to be estimated for each frequency bin. For example, for SDW-MWF [5], having M microphones per hearing aid and using an FFT of length L, it is necessary to estimate \( L/2 \) correlation matrices of size \( 2M \times 2M \) for the signal and noise, and to solve the same number of linear system of equations. Second, the processing delay introduced by the MWF processing depends on L. Since this delay is crucial for a digital hearing aid, it is necessary to keep L as small as possible. Although both computational resources and processing delay may be reduced with the use of smaller FFT lengths, Docto et al. [3] have showed that the performance of the MWF methods is degraded when L is decreased.

To reduce computational resources and processing delay, this paper proposes the use an auditory processing rather than an FFT processing. An FFT-based processing can be seen as an uniform filter bank (Fig. 1). In this case, increasing L increases the number of filters, reduces the bandwidth of each filter, and then, improves the frequency resolution at both low and high frequencies. However, it is widely known that the frequency response of the human auditory system is more selective at low frequencies (Fig. 1), and this low-frequency range is more relevant for the improvement of speech intelligibility. Therefore, in an FFT-based processing, to reach the low-frequency resolution of the human auditory system, it is necessary to employ a large L. Using larger L introduces more filters in the high-frequency range, but improving the SNR at these high-frequency bins does not contribute significantly to the speech intelligibility. Therefore, it is possible to reduce the computational cost without degrading the performance in a MWF method by using an auditory representation instead of an FFT representation. The additional advantage of this approach is that the number of sub-bands is fixed, and their respective bandwidths are independent on the frame length L. A typical number of auditory sub-bands is 24 for a sampling frequency of 22 kHz. Therefore, the computational cost savings achieved with this processing are very significant. This paper shows that the use of this auditory representation reduces the computational cost, reduces the processing delay, and achieves significant SNR improvements. In addition, this paper shows that the existing MWF techniques proposed in the FFT domain are still valid in this auditory domain.

2. FFT-based MWF

In the FFT domain, for a particular frequency bin \( f \), the signals received by all microphones can be described by the vector \( \mathbf{y}(f) = [y_1(f), y_2(f), ..., y_{2M}(f)]^T \), where \( M \) is the number of microphones for each hearing aid. Assuming an additive background noise, the vector \( \mathbf{y}(f) \) can be expressed as \( \mathbf{y}(f) = \mathbf{x}(f) + \mathbf{v}(f) \), where \( \mathbf{x}(f) \) and \( \mathbf{v}(f) \) are the vectors that describe the target signal and noise components at the frequency bin \( f \).
Figure 1: Filter-banks used in an FFT of length \( L = 128 \) and sampling frequency 22kHz (top) and a wavelet packet tailored to properties of the human auditory system (bottom).

In the MWF framework, the filter coefficients are computed by the minimization of the minimum mean square error (MMSE) between the filter output and the desired signal. There are different MWF objective functions proposed in the literature [1, 3, 5]. In this paper, the most widely-known MWF method, called speech distortion weighted multichannel Wiener filter (SDW-MWF) [5], is used. In SDW-MWF, the weights at the right and left channel are computed by the minimization of

\[
J_{SDW}(W) = \mathbb{E} \left\{ \left\| (e_L - W_L)X + (e_R - W_R)X \right\|^2 + \mu \left\| W_H^H V W_H W \right\|^2 \right\}
\]

where \( \mu \) denotes a trade-off parameter between noise reduction and speech distortion; \( e_L \) and \( e_R \) are unitary vectors of length \( 2M \) describing the position of the reference microphones for the left and right hearing aid. The frequency index \( f \) was dropped from \( X, Y, W_L, \) and \( W_R \) for mathematical convenience. After minimization, the filter coefficients are given by

\[
W_L(f) = (R_L(f) + \mu R_s(f))^{-1} R_s(f) e_L
\]

\[
W_R(f) = (R_L(f) + \mu R_s(f))^{-1} R_s(f) e_R
\]

where \( R_L \) and \( R_s \) are the second-order statistics describing the speech and noise correlation matrices, defined as \( R_L(f) = \mathbb{E} \{ X(f)X(f)^H \} \) and \( R_s(f) = \mathbb{E} \{ V(f)V(f)^H \} \).

For practical implementations, the correlation matrix \( R_s \) can be estimated during the unvoiced segments, using a voice activity detector (VAD). Under the assumption of statistical independence of the target and noise signals, the correlation matrix \( R_s \) can be estimated as \( R_s = R_{sv} - R_{vv} \), where \( R_{sv} \) is estimated during voice (or speech) segments.

### 3. Proposed method

The idea behind the proposed method is to replace the FFT processing by a processing that resembles the auditory filter bank (Fig. 2). In the proposed method, the signals received at each microphone \( y_i(n), i = 1, \ldots, 2M \), are passed through an auditory filter bank, which decomposes each input signal into \( K \) sub-bands. In this case, each sub-band output is represented by \( y_{i,k}(n') \), where \( k \) corresponds to the sub-band index, and \( n' \) is a time-index in the auditory domain. If the operation applied by the auditory filter bank is linear, it is possible to describe \( y_{i,k} \) as \( y_{i,k} = x_{i,k} + v_{i,k} \), where \( x \) and \( v \) correspond to the target and noise components in the \( k \)-th sub-band and \( i \)-th microphone, respectively. The filter-bank outputs can be described more conveniently in the form \( Y_k = [y_{1,k} \ y_{2,k} \cdots \ y_{2M,k}]^T \), where \( Y_k \) is a vector of length \( 2M \) holding the information of the \( k \)-th sub-band at the time-index \( n' \). Again, \( Y_k = X_k + V_k \). The output of the Wiener filter for the left channel, in the auditory domain, is given by \( z_{L,k} = W_{L,k}^T Y_k \).

Under the assumption of linearity of the transform used for the auditory representation, the objective function in (1) is still valid, and so the equations to estimate the filter coefficients given by (2) and (3). In this case, the correlation matrices are defined for each sub-band instead of each frequency bin, i.e.

\[
R_{v,k} = \mathbb{E} \{ X_k X_k^H \} \quad \text{and} \quad R_s(f) = \mathbb{E} \{ V(f)V(f)^H \}.
\]

This suggests that any MWF method proposed in the FFT domain is still valid for the auditory domain.

There are two possible candidates to perform the transformation to the auditory domain: an IIR filter bank and a wavelet packet (WP). The IIR filter bank has the advantage of providing a very small processing delay. Although the statistics and weights can be updated at each time index, the computational cost involved in this approach is very expensive compared to an FFT-based processing. Even though these statistics can be updated in a frame-by-frame basis, the total number of operations is still high compared to an FFT-based processing. In addition, the IIR filter bank may not provide a perfect reconstruction as in a WP.

On the other hand, in a WP approach, it is possible to reduce the computational cost and processing delay, and ensure the absence of block-processing artifacts. If the input frame length is \( L \), the number of samples, \( L_k \), at each sub-band is a power-of-two fraction of \( L \). Hence, the estimation of the statistics for each sub-band involves a lower number of operations compared to the IIR filter-bank case. There are different WP trees proposed in the literature to imitate the behavior of the human auditory system. Karmakar et al. [6] proposed a WP tree that resembles the auditory filter bank, which was designed to satisfy perceptual criteria. The frequency response of this filter bank is shown in the Fig. 1.

The estimation of the statistics in the proposed method is performed every \( L \) samples using a first-order estimator:

\[
R_{q,k}(l) = \alpha R_{q,k}(l) + \frac{(1 - \alpha)}{L_k} Y_k(l)Y_k^T(l)
\]

where \( q = x \) or \( v, \alpha \) is a time constant, \( l \) is the frame index, and \( Y_k(l) \) is a matrix of size \( 2M \times L_k \). Each \( i \)-th row of \( Y_k \) holds the output vector of the \( k \)-th sub-band taken from the WP of the \( i \)-th input microphone.

### 4. Experimental setup

The performance of the proposed method (SDW-PMWF) and the existing method based on FFT processing (SDW-MWF) [5] was tested under two scenarios: diffusive noise and babble noise. To construct the mixtures of each scenario, the target signal was filtered with the head related transfer functions.

Figure 2: Proposed processing using auditory filter-banks and MWF framework for binaural noise reduction.
(HRTF) measured for a KEMAR manikin [7]. To simulate the use of two microphones per each hearing device, an uniform linear array with a separation of 8mm was used. The target signal (speech recordings of ten different speakers and sentences) was placed at eight different azimuth angles: 0°, 30°, 90°, 120°, 180°, 240°, 270° and 330°, where 0° corresponds to the front of the KEMAR, 90° corresponds to the right ear, and 270° to the left ear. For all scenarios, the interfering signals were added to the target signal at different SNR. For diffusive noise scenario, different uncorrelated pink noise sources were played simultaneously at 18 different spatial locations. For the babble noise scenario, the noise source was recorded in a cafeteria and added to the speech samples processed with the HRTFs. For all mixtures a sampling frequency of $f_s = 22$ kHz was used.

The methods were implemented using block-processing in order to simulate a real-time implementation. To obtain the upper-bound performance, the statistics for the signal and noise were estimated using a perfect voice activity detector (VAD). Mother wavelets Daubechies 4 (db4) and 8 (db8) were used for the WP. These mother wavelets have been shown in the literature to provide good performance for speech applications.

The performance of these techniques was analyzed using two metrics, the broadband intelligibility weighted SNR improvement [8] ($\Delta$SNR), and the objective quality assessment measure (PEMO-Q) [9]. $\Delta$SNR values reported in this paper corresponds to the average over all target speakers and angles.

5. Results and discussion

An important property to explore in the proposed processing is the identification of trade-offs between the SNR improvement ($\Delta$SNR) and sound quality (PEMO-Q) with respect to the frame length $L$ and the parameter $\mu$. Figs. 3 and 4 show the SNR improvement for the diffusive and babble noise scenarios as a function of $L$ and $\mu$, and Figure 5, the sound quality for a babble noise scenario as function of $L$ and $\mu$. It is clear from these figures that SNR improvement is increased with the use of a large $\mu$, as expected, and a large $\mu$ also provides better sound quality. The latter can be explained by the fact that increasing $\mu$ provides a higher noise reduction, and then it improves the output sound quality. It is also important to remark that the proposed processing requires a small frame length to achieve higher SNR improvement and sound quality. This re-

![Figure 3: SNR improvement for diffusive noise scenario vs the frame length $L$ and parameter $\mu$. Input SNR = 0dB. WP: db4.](image)

![Figure 4: SNR improvement for babble noise scenario vs the frame length $L$ and parameter $\mu$. Input SNR = 0dB. WP: db4.](image)

![Figure 5: Sound quality for babble noise scenario vs the frame length $L$ and parameter $\mu$. Input SNR = 0dB. A PEMO-Q value of 1 corresponds to a clean signal. WP: db4.](image)

![Figure 6: SNR improvement for diffusive noise scenario at different input SNR.](image)

![Figure 7: SNR improvement for babble noise scenario at different input SNR.](image)

![Figure 8: Sound quality for babble noise scenario at different input SNR. A PEMO-Q value of 1 corresponds to a clean signal. WP: db4.](image)
sult is particularly important for the real-time implementation in a digital hearing aids since using a smaller \( L \) involves a reduction in the processing delay. In contrast, Doclo et al. [3] identified that for a FFT-based processing, \( L \) needs to be high to achieve good SNR improvement. The fact that the performance in the proposed processing is increased with the use of smaller \( L \) can be explained by the independence of the number of sub-bands on the frame length \( L \), and the way how the statistics are updated. In FFT-based processing, the low-frequency resolution depends on the frame length \( L \), but in the proposed approach the number of sub-bands is fixed and independent on \( L \), so that the frequency resolution is expected to be high at low frequencies regardless the value of \( L \). In MWF methods, the performance is highly dependent on the estimation of the signal and noise statistics. Therefore, when \( L \) is small, the estimators are able to track rapid changes in speech and noise statistics, providing better estimation, and then better noise reduction.

For comparison purposes, the SNR improvement of the proposed processing (SDW-PMWF) is compared to the FFT-based processing (SDW-MWF) under diffusive and babble noise scenarios at different input SNR (Figs. 6 and 7). The frame length for both methods was chosen to be \( L = 128 \) in order to establish a comparison of these methods under the same processing delay. The proposed processing uses the optimal parameter selected from the previous analysis, i.e. \( \mu = 20 \), and FFT-based processing uses \( \mu = 5 \). The value of \( \mu \) for SDW-MWF was selected according to the results presented in [3]. From these figures, it is evident that the proposed processing provides a significant SNR improvement over the FFT-based processing of a least 4 – 5dB and 1 – 3dB for diffusive and babble noise scenarios, respectively. In addition, this significant \( \Delta \text{SNR} \) also provides an improvement in the output sound quality (Fig. 8). Moreover, the performance of the proposed method depends on the mother wavelet. \( \Delta \text{SNR} \) is reduced less than 1dB when a db8 is used instead of a db4. Although db4 offers higher SNR improvement, the speech quality provided by this WP is slightly lower than db8.

Another important goal of the proposed method is the reduction of the computational complexity. Fig. 9 shows the number of operations required by the proposed and the FFT-based processing to process an input sample at 22 kHz sampling rate, i.e. the total number of operations required to process a frame of length \( L \) is divided into the frame length. These number of operations is discriminated into multiplications (MPY), additions (ADD), and divisions (DIV). This figure shows how the number of operations in an FFT-based processing is reduced dramatically (around 70%) by using the proposed processing with db4.

Finally, neglecting the delay introduced by the transceiver, the processing delay of the proposed algorithm depends exclusively on the frame length \( L \). Since \( L = 128 \) provides good SNR improvement and quality, the processing delay is lower than 6 ms at 22 kHz sampling rate. On the other hand, the signal delay depends on the wavelet filter length. Since these filter lengths are 8 and 16 for db4 and db8, respectively, the signal delay is less than 1 ms at 22 kHz sampling rate.

6. Conclusions
This paper proposed the use of a wavelet packet (WP) to substitute the FFT processing in a Multichannel Wiener filter (MWF) method. The WP tree used in this paper resembles the auditory filter bank, providing higher low-frequency resolution than an FFT. The use of this WP ensures that the mathematical frame

![Figure 9: Computational cost of the proposed and FFT-based processing for different frame lengths.](image)

work derived originally in the FFT domain is still valid. This processing provides different advantages with respect to the FFT-based processing. First, the SNR improvement and output sound quality are higher in the proposed processing. Second, the processing delay can be reduced without degrading the performance of the algorithm. Third, the computational cost of the propose processing is lower than FFT-based processing.

7. References