Wavelet Transform Based Coherence Function For Multi-Channel Speech Enhancement

Djamila Mahmoudi and Andrzej Drygajlo
Signal Processing Laboratory, Swiss Federal Institute of Technology at Lausanne,
CH-1015 Lausanne, SWITZERLAND,
e-mail: {Djamila.Mahmoudi,Andrzej Drygajlo}@epfl.ch

ABSTRACT

This paper addresses the problem of enhancing a speech signal acquired by a microphone array used for hands-free voice communication applications. A new algorithm based on the coherence function developed in the wavelet domain and applied to the beamforming output signal is presented. The wavelet coherence function based nonlinear post-filtering provides a further noise suppression. Its performances is comparable to that of the wavelet Wiener filter proposed in our earlier work [1].

1 INTRODUCTION

Most of the hands-free voice communication systems are intended to work in real conditions where interfering sources like noise and reverberation are are present. Many speech enhancement algorithms have been proposed in order to provide a good quality of communication. For hands-free applications, microphone array systems have shown their superiority over the one channel systems thanks to its independence from the nature of the interfering signals and its capability to automatically alter its direction of reception. In fact, this property makes it suitable for situations involving a moving source.

One of the commonly used microphone array speech enhancement methods is conventional delay-and-sum beamforming with an additional post-filtering as proposed in [2, 3]. This method has proven its efficiency for noise reduction. However, important distortions are introduced by the post-filter, particularly because of the inaccurate speech and noise power spectral densities (psd) estimates. Consequently, smoothing techniques for improving the psd estimates, have been proposed. Furthermore, poor performance is obtained at low frequencies because of a high spatial coherence between noises in the microphone signals. This is due to the small spacing required between microphones in the uniform array.

As a speech acquisition system, the microphone array is intended for voice communication application through a terminal. Consequently, the wideband nature of the speech signal and the physical limitations of the communication terminal result in some constraints in the microphone array design.

A sub-array sub-band decomposition has been proposed to ensure a sufficient directivity, without spatial aliasing, over the whole speech bandwidth. It uses a small array with a logarithmic distribution of microphones to cover the telephony band of the speech signal. An octave-band decomposition that follows a dyadic decomposition of the frequency axis has been proposed [4]. In fact, it is similar to the one suggested for the harmonically nested arrays used for wideband signals [5, 6]. However, it is well known that the structure of computations in a discrete wavelet transform and in an octave-band filter bank are identical. Thus, the wavelet transform finds an immediate application in microphone array speech enhancement systems. This transform offers many advantages, such as perfect reconstruction and considerable computational savings due to the critical sub-sampling. Furthermore, it is well suited to the non-stationary speech signal, and a fast algorithm for its calculation is also available.

Extensions of the simple conventional beamforming to overcome its inherent limitations have been proposed by many authors [2, 3]. One approach, commonly used in speech enhancement, is to pass the beamforming signal through a post-filter in order to obtain further noise suppression. For this purpose, a Wiener filter expressed in the wavelet domain was proposed in our earlier work [1]. The basic idea of the proposed system is that both the sub-array sub-band beamforming and the post-filtering are performed in the wavelet domain as illustrated in Fig. 1. Good performance is thus obtained with this algorithm thanks to the multi-resolution property.

In this paper, we propose a new approach based on a the coherence function estimated in the wavelet domain. This method exploits the behaviour of the coherence function to provide a new procedure for the post-filtering in microphone array speech enhancement. Hence, the contribution presented here differs from that in [7, 8] by considering a coherence function calculated between the beamforming signal and the reference mi-

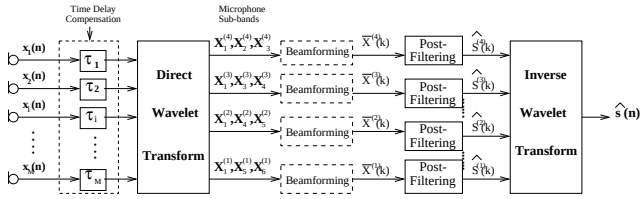

Figure 1: *Block diagram of noise reduction algorithm.*

crophone signal. Indeed, our coherence filtering is proposed as a nonlinear post-filter applied to the beamforming output in order to improve it.

2 PROBLEM FORMULATION

The coherence function is commonly used to measure the similarity between two signals. Coherence based filtering was first proposed by Allen et al. for speech dereverberation [9]. It was extended to noise reduction by Kaneda and Tohyama [7], and Le Bouquin and Faucon [8] and it was applied to two-channel speech acquisition systems. In these works, two microphones are used where the second offers an additional version of the noisy signal that permit to estimate noise, especially when noise is nonstationary. The microphone signals $x_1(n)$ and $x_2(n)$ are modeled as:

$$\begin{cases} x_1(n) &= s_1(n) + n_1(n) \\ x_2(n) &= s_2(n) + n_2(n), \end{cases} \tag{1}$$

where $s_1(n)$ and $s_2(n)$ are two highly correlated versions of the clean signal $s(n)$, and $n_1(n)$ and $n_2(n)$ are the uncorrelated noises that affect the clean signal $s(n)$. The coherence function calculated between $x_1(n)$ and $x_2(n)$ is

$$, _{x_1x_2}(\omega) = \frac{\Phi_{x_1x_2}(\omega)}{\sqrt{\Phi_{x_1x_1}(\omega)\Phi_{x_2x_2,p}(\omega)}}, \tag{2}$$

where $\Phi_{x_ix_i}(\omega)$ and $\Phi_{x_ix_j}(\omega)$, $i,j = 1,2$ are the auto-power and cross-power density spectra of the signals $x_i(n)$ and $x_j(n)$ respectively. The magnitude squared coherence defined as $MSC(\omega) = |, _{x_1x_2}(\omega)|^2$ provides revelant and significant information about the presence or the absence of the speech signal: if $MSC(\omega)$ is close to one, speech is present and if $MSC(\omega)$ is close to zero, speech is absent. In the latter case, the current frame contains only noise (pause) which can be removed using a nonlinear filter. Unfortunately, a processing which is based on the model given in (1) as done in [7, 8] is far from being exploitable in our post-filtering task. The reason is that noises captured at two microphones are strongly correlated because of the small microphone spacing required to avoid spatial aliasing in the speech bandwidth.

In our work, the basic idea of the use of the coherence function in the post-filtering for microphone arrays consists in exploiting the change in the additive noise structure as a result of conventional beamforming. In other terms, once the time delay is compensated for all the microphone signals according to the desired source direction and the beamforming operation is performed yielding $y(n) = \bar{x}(n)$, the noise components are also added and averaged while they are unphased. This results in a beamforming noise $\bar{n}(n)$, which is statistically different from the noises $n_i(n)$ at each microphone. Consequently, the assumption of the non-correlated noises is verified. Furthermore, the beamforming signal $y(n)$ contains the same components $s(n)$ as the reference microphone signals. Thus, the coherence function will be calculated between the beamforming signal $\bar{x}(n)$ and the reference microphone signal $x_1(n)$.

3 WAVELET COHERENCE FUNCTION

Let $\mathbf{V}$ and $\mathbf{V}^{-1}$ denote the matrices representing the discrete wavelet transform (DWT) and its inverse respectively. $\mathbf{V}^{-1} = \mathbf{V}^T$ because $\mathbf{V}$ is orthogonal. DWT performed on the discrete-time signal $\mathbf{x}$ gives

$$\mathbf{X} = \mathbf{V} \cdot \mathbf{x} = \mathbf{S} + \mathbf{N}, \quad \text{and} \quad \mathbf{x} = \mathbf{V}^{-1} \cdot \mathbf{X} \tag{3}$$

the capital letters symbolize the vectors containing the wavelet coefficients of the signals noted by $X^{(m)}(k)$; m and k are the scaling and shift parameters respectively.

The wavelet coherence function, which is analogous to the FFT-based coherence function is introduced to estimate the degree of similarity between any two signals in the time-frequency domain. Let $\beta \in [0.1, 0.5]$ be a forgetting factor. We define the wavelet coherence function between two transformed signals $\mathbf{X}$ and $\mathbf{Y}$ as

$$, \overset{(m)}{_{\mathbf{XY},p}}(k) = \frac{\Phi^{(m)}_{\mathbf{XY},p}(k)}{\sqrt{\Phi^{(m)}_{\mathbf{XX},p}(k)\Phi^{(m)}_{\mathbf{YY},p}(k)}}, \tag{4}$$

where $\Phi^{(m)}_{\mathbf{XY},p}$ is the wavelet auto-power spectrum if $\mathbf{X} = \mathbf{Y}$ and the wavelet cross-power spectrum elsewhere, of the current frame indexed by p. It is calculated as follows:

$$\Phi^{(m)}_{\mathbf{XY},p}(k) = \beta \cdot \Phi^{(m)}_{\mathbf{XY},p-1}(k) + X^{(m)}(k) \cdot Y^{(m)}(k), \tag{5}$$

Note that each frame contains P discrete-time samples.

4 POST-FILTERING WITH WAVELET COHERENCE FUNCTION

In this work, a novel post-filtering based on the wavelet coherence function is proposed. The detailed structure of the proposed speech enhancement system is illustrated in Fig. 2. The microphone signals are time-delay compensated and then transformed to the time-frequency domain. Uniform sub-arrays that operate on their appropriate sub-bands are formed [4]. The array consists of 6 microphones distributed logarithmically and each microphone signal is decomposed into 4 octave sub-bands. Let M_s be the number of microphones

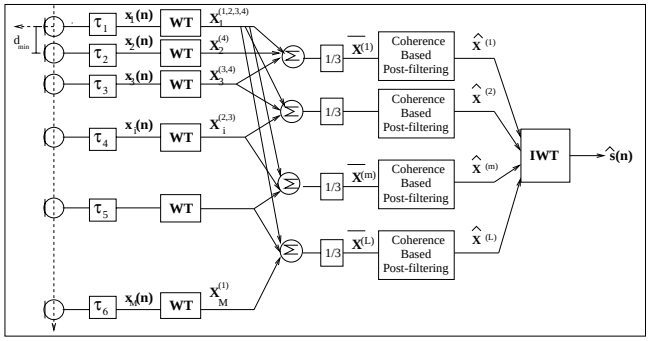

Figure 2: *Block diagram of the overall system.*

in each sub-array, the conventional beamforming is performed in the sub-bands using their corresponding sub-arrays yielding $Y^{(m)}(k) = \bar{X}^{(m)}(k)$. In other terms, the components of this vector are summed and averaged to form the sub-band beamforming coefficients $Y^{(m)}(k)$ at mth stage as follows:

$$Y^{(m)}(k) = \frac{1}{M_s} \sum_{i=1}^{M_s} X_i^{(m)}(k) = S^{(m)}(k) + \bar{N}^{(m)}(k). \quad (6)$$

Then, the sub-band beamforming signal $Y^{(m)}(k)$ passes through a nonlinear post-filter H to increase the noise reduction performance. Note that $\mathbf{X}^{(m)}$ is the vector of the wavelet coefficients that correspond to the mth sub-band if k is omitted.

Following the motivation developed in Sec. 2, our post-filter which is applied to the beamforming signal is based on the coherence function calculated between the beamforming signal and the signal of the first microphone (taken as reference microphone). These signals are modelled as.

$$\begin{cases} X_1^{(m)}(k) &= S^{(m)}(k) + N^{(m)}(k) \\ \bar{X}^{(m)}(k) &= S^{(m)}(k) + \bar{N}^{(m)}(k), \end{cases} \quad (7)$$

By analogy to the MSC calculated in the Fourier domain and using the definition given in Eq. (4), we define a wavelet coherence function of order σ as

$$C_{\mathbf{XY},p}^{(m)}(k) = \left[, \, {}^{(m)}_{\mathbf{XY},p}(k) \right]^{\sigma} \qquad \sigma \in \mathcal{R}^+, \quad (8)$$

Let $G_p^{(m)}$ be the averaged coherence over the whole mth sub-band for the current speech frame. At the scale m, each frame consists of P_m wavelet coefficients where $P_m = P/2^m$. Note that $G_p^{(m)}$ provides one value per sub-band since it is expressed as

$$G_p^{(m)} = \frac{1}{P_m} \sum_{k=1}^{P_m} C_{\mathbf{X}_1 \bar{\mathbf{X}},p}(k), \quad (9)$$

Considering the model given in Eq. (7) and taking into account the weaker correlation between the noise components $N_1^{(m)}(k)$ and $\bar{N}^{(m)}(k)$, $G_p^{(m)}$ provides revelant and significant information about the presence or absence of the speech. Fig. 3 shows the $G_p^{(m)}$ behaviour for $\sigma = 2$. It generally exhibits a significant change of value when the speech energy is absent. This important property leads to coherence coefficients which are small during pauses and close to one when the speech signal is present. In fact, it has been observed that the coherence function calculated in the wavelet domain leads to very low values in pauses compared to the FFT-based coherence performed for the same conditions.

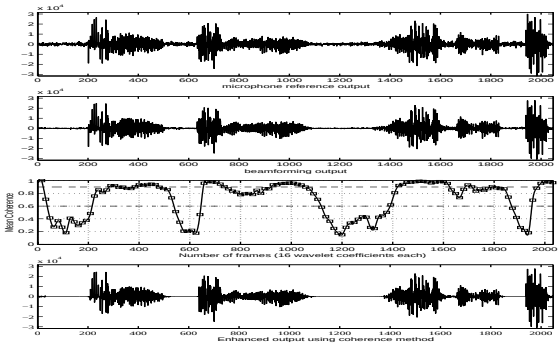

Figure 3: *Wavelet noise reduction using coherence strategy (sub-band: [0.5 kHz,1 kHz]).*

4.1 Noise Reduction Rules

The wavelet coherence function is exploited in a nonlinear post-filter H to improve the sub-band beamforming signal as shown in Fig. 4. The speech enhancement based on post-filtering is achieved using the following noise reduction rules. They are derived by observing the behaviour of $G_p^{(m)}$ in the different sub-bands.

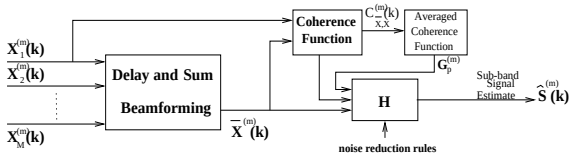

Figure 4: *Block diagram of the proposed algorithm.*

$$\hat{S}^{(m)}(k) = \begin{cases} \bar{X}^{(m)}(k), & \text{if} \quad G_p^{(m)} > S_{max} \\ (G_p^{(m)})^{\sigma} \cdot \bar{X}^{(m)}(k) & \text{if} \quad G_p^{(m)} \leq S_{min} \\ C_{\mathbf{X}_1 \bar{\mathbf{X}},p}^{(m)}(k) \cdot G_p^{(m)} \cdot \bar{X}^{(m)}(k) & \text{elsewhere.} \end{cases} \quad (10)$$

The choice of the thresholds is important for selecting the frames of the signal considered as noisy and the speech parts which will be remain untreated. The thresholds S_{min} and S_{max} are chosen empirically, based on listening judgements. They are fixed and chosen equal to 0.6 and 0.9 respectively.

5 PERFORMANCE EVALUATION

The proposed algorithm was tested using a simulated non-symmetric array with 6 microphones distributed logarithmically where the smallest spacing $d_{min} = 5cm$. For this procedure, we have adopted the Daubechies' prototype filters of 3rd order. It has been seen that the 3rd order is sufficient for the speech frame to be enhanced. For the performance evaluation, we have chosen the segmental SNR (SNR_{seg}). Fig. 5 compares the resulting signals from two post-filters: Wiener filter and coherence filter for $\sigma = 2$. Fig. 6 shows the achieved

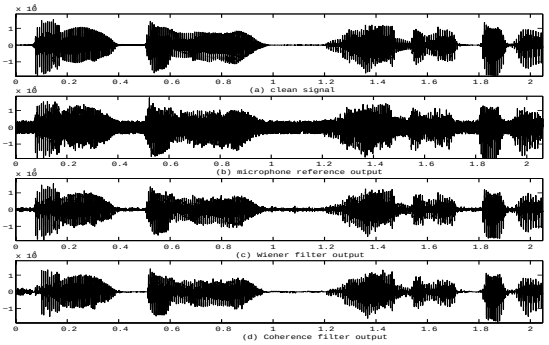

Figure 5: *Signals using the proposed algorithm.*

noise reduction factor for the different parts of the signal. It can be seen that the proposed algorithm provides good results compared to those obtained with wavelet transform based Wiener filter proposed in [1], especially during pauses. Furthermore, unlike the already proposed noise reduction methods based on FFT coherence function [10], informal listening tests showed that the resulting signal sounds natural and no discontinuities are noticed.

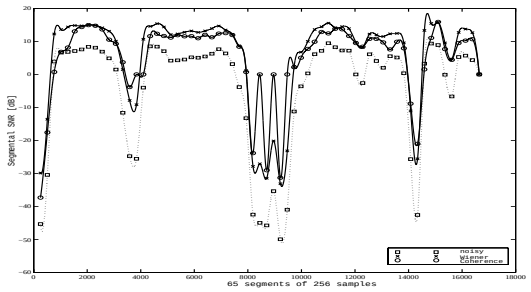

Figure 6: *Achieved noise reduction in terms of* SNR_{seg}.

The proposed algorithm is thus suitable for an efficient speech enhancement system, particularly for moderate SNR. Indeed, noise still present in the beamforming signal is audible only during the unvoiced parts and pauses of the signal.

6 CONCLUSION

In this paper, a new noise reduction procedure based on the coherence function performed in the wavelet domain is introduced as a nonlinear post-filter to separately enhance the sub-band beamforming signals. The new algorithm provides a significant noise suppression with negligble distortions, especially during pauses. This is confirmed by the objective measures and the informal listening tests. The proposed method is thus suitable for an efficient speech enhancement system, particularly for moderate SNR where the noise present in the beamforming signal is more audible during the unvoiced parts and pauses of the signal. Ongoing work consists in exploiting the behaviour of the wavelet coherence function with different values of σ to increase the noise reduction system performance.

References

[1] D. Mahmoudi, "A Microphone Array for Speech Enhancement using Multiresolution Wavelet Transform", *in Proc. of Eurospeech'97*, pp. 339–342, Rhodes, Greece, September 1997.

[2] R. Zelinski, "A Micropohone Array with Adaptive Postfiltering for Noise Reduction in Reverberant Rooms", *in Proc. of ICASSP'88*, pp. 2578–2581, 1988.

[3] K. U. Simmer and A. Wasiljeff, "Adaptive Microphone Arrays for Noise Suppression in the Frequency Domain", *in* 2^{nd} *Cost 229 Workshop on Adap. Algo. in Com., France*, pp. 185–194, Sep. 1992.

[4] D. Mahmoudi and A. Drygajlo, "Multiresolution Microphone Array for Speech Source Acquisition and Tracking", *in Proc. of IWAENC'97*, pp. 116–119, London, UK, September 1997.

[5] W. Kellermann, "Self-Steering Digital Microphone Array", *in Proc. of ICASSP'91*, vol. 4, pp. 3581–3584, July 1991.

[6] Y. Mahieux, A. Gilloire, and G. Le Tourneur, "A Microphone Array for Multimedia Applications", *in IEEE Workshop on App. of Sig. Proc. to Aud. and Acous.*, pp. 529–532, New York, USA, October 1993.

[7] Y. Kaneda and M. Tohyama, "Noise Suppression Signal Processing using 2-point Received Signals", *Electronics and communications*, vol. 67A, pp. 19–28, April 1984.

[8] R. Le Bouquin and G. Faucon, "Using the Coherence Function for Noise Reduction", *IEE Proceedings*, vol. 139, pp. 276–280, June 1992.

[9] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone Signal Processing Technique to Remove Room Reverberation from Speech Signals", *J. Acoust. Soc. Am.*, vol. 62, pp. 912–915, 1977.

[10] B. Le Bouquin-Jeannes, A. A. Azirani, and G. Faucon, "Enhancement of Speech Degraded by Coherent and Incoherent Noise Using a Cross-Spectral Estimator", *IEEE Trans. on Speech and Audio Proc.*, vol. 5, pp. 484–487, September 1997.