Analysis of the Evolution Speech Enhancement Methods in Wavelet Domain

**Caio C. E. de Abreu**  
Department of Electrical Engineering, FEIS - UNESP  
15385-000, Ilha Solteira, SP  
E-mail: caioenside@aluno.feis.unesp.br

**Marco A. Q. Duarte**  
Department of Mathematics - UEMS  
79540-000, Cassilândia, MS  
E-mail: marco@uems.br

**Francisco Villarreal**  
Department of Mathematics, FEIS – UNESP  
15385-000, Ilha Solteira, SP  
E-mail: villa@mat.feis.unesp.br

**Abstract**: This paper discusses advances in speech enhancement methods based on wavelets. The work is based on an updated research about the main speech enhancement methods that use or benefit some form of wavelet analysis. The problem is studied as an evolutionary process of interdisciplinarity. It was found that the research are geared on look for new techniques of data analysis in order to take advantage of the speech signal features extracted from discrete wavelet transform.

**keywords**: Speech enhancement, wavelet de-noising, analysis of methods

1 **Introduction**

The improvement of the wavelet analysis, specifically the discrete wavelet transform (DWT), enabled the development of new techniques for digital signal processing in several areas. Among them, there are the techniques geared for speech enhancement, picture enhancement and applications in medicine. Speech enhancement is essential in many branches of telecommunications and of the entertainment industry. Examples are the applications in signal encoding and automatic speech recognition.

The problem of speech enhancement consists in recover the original signal from the corrupted speech signal. In other words, it consists in recover $x[n]$ from equation (1).

$$y[n] = x[n] + w[n],$$

where $w[n]$ is the additive noise. Classical speech enhancement methods can be generally divided into two groups, Fast Fourier Transform (FFT) based filtering [2, 14, 20] and wavelet thresholding [6, 7, 10, 16]. In FFT based filtering methods, the frequency spectrum of the noisy speech is modified to reduce the noise. These methods depend on good noise estimate for then perform Spectral Subtraction. Originally proposed by [7], wavelet thresholding accepted as noisy coefficients those ones with absolute value below of a certain value and then those coefficients are modified.

This paper presents an updated research about the main speech enhancement methods based on wavelets in order to check the tendencies for future research in the area.
2 Discrete wavelet transform - DWT

An efficient algorithm for the computation of DWT was proposed by Mallat [12]. This method, called fast wavelet transform, uses a digital filter bank in a tree structure. The coefficients of low-pass and high-pass filters, used in the decomposition process, are generated from a wavelet function chosen a priori. For each wavelet decomposition level, approximation and detail coefficients are acquired. In the transition from one wavelet level to another, the decomposition is applied again, however, only on the approximation coefficients.

The decomposition level is associated with the resolution level of the DWT. The higher the decomposition level the greater will be the resolution level, capturing more signal detail and minimizing possible loss of the information. Thus, the DWT is a powerful tool for studying nonstationary signals [4, 8]. At the highest wavelet decomposition levels, the energy of the signal is concentrated in a small number of coefficients [4, 19]. This fact accepts different interpretations and enables its interdisciplinarity.

3 Evolution of speech enhancement methods based in wavelets

When applied on a speech signal, the DWT extracts coherent structures, and this information is expressed by wavelet coefficients in the transformed domain. If correctly interpreted, the DWT becomes a powerful tool to the speech enhancement problem. Some characteristics are exploited by wavelet methods: The possibility of predetermined setting of resolution level during decomposition process, concentration of the energy of the transformed signal into a small number of coefficients and the achievement of different frequency sub-bands for the decomposed signal. The last characteristic is the most explored, making possible to build systematized silence/voice activity detectors and the detection of singularities in a corrupted speech signal [8]. In [21], the authors developed a system for the detection of impulsive colored noise in time domain (kind of singularity) in the corrupted speech signal. For this reason, the authors analyzed differences among energy distribution of corrupted signal and impulse noise by frequency sub-bands provided by DWT (See Figure 1, where the coefficients of low frequency are located in the frequency band L).

![Figure 1: Frequency sub-bands provided by DWT.](image-url)

The earliest methods of speech enhancement in the wavelet domain were based on thresholding [8]. Those methods took advantage of the concentration of the speech energy on low frequency bands [5]. Using that fact, together with wavelet analysis, which provides different frequency sub-bands for a corrupted signal, a threshold value was estimated. Thus, assuming as noisy coefficients those whose absolute values were under the threshold value. The main job of those methods consisted in the estimation of the threshold value [8, 15]. But, according to [17], in spite of the presence good results for threshold methods in the literature, those methods could confuse voice with noise due to the fact that some voice coefficients have absolute value below the threshold calculated, generating discomfort to the listener.

In an evolutionary process, trying to solve the problems presented by thresholding methods, some methods that act on the corrupted signal according to itself content were proposed [1, 18]. Considered nonthreshold, these methods overcome their predecessors in requisite psychouacoustic quality of enhanced speech. This is due to a uniform reduction noise
along throughout the signal, avoiding sounds inconvenient generated by search of the best
threshold, being more enjoyable to the listener [18]. It does not exist in the literature many
speech enhancement methods in the wavelet domain considered nontreshold. This is due to
the difficulty in developing methods with these characteristics and a recently methodology.

The abundance of details provided by wavelet analysis has attracted the attention of
researchers from various fields. Particularly, for the case in speech enhancement, emerged
innovative studies that use of other data analysis techniques to highlight the properties of the
DWT [3]. In [15], the authors propose a method to calculate the threshold based on the
statistical modeling of the Teager operator. Applied on wavelet packet coefficients of de noisy
speech, the Teager operator extracts the signal energy based on mechanical and physical
considerations [15]. In this sense, [11] combine the DWT with two other tools and propose a
new method for speech enhancement. It is observed in this work which the authors combine
the Hilbert-Huang transform, the empirical mode decomposition (EMD) and a DWT with
purpose of building a wavelet filter based on thresholding.

As noted in [3, 11, 15] , the research is being directed to the improvement of
estimation of the threshold by means of other techniques to analyze data. Furthermore,
emphasizing the interdisciplinarity of DWT, some methods developed recently proposed the
junction of two important techniques in speech enhancement: spectral subtraction and wavelet
denoising [9, 13]. In [9], it was proposed a noise reduction method based on spectral
subtraction performed in the wavelet domain. The method consists in the application of the
traditional spectral subtraction technique on approximation and detail wavelet coefficients.
Similarly, in [13] the authors used the concept of frequency sub-bands in the wavelet domain
together with spectral subtraction techniques. The idea consisted in applying the spectral
subtraction techniques to reduce noise in two low frequency sub-bands, after that, thresholding
was applied in the other sub-bands. Thereby, it avoided the degradation of the low frequency
coefficients, which can be more easily corrupted when using some kind of threshold, improving
the quality of the enhanced signal. Taking advantage of the main features inherent to both
methods, the authors combined a strong noise reduction with a good preservation of the voice
coefficients.

4. Computational analysis of evolution speech enhancement based in wavelets

For purposes of comparison, in this section four different speech enhancement methods will be
implemented; a thresholding method (Sigmoidal thresholding) [8, 17], two nontresholding
methods (A and B) proposed in [1] and [18], respectively, and a method which combines
spectral subtraction witch DWT (FFT/Wavelet) proposed in [9].

In [17], the authors evaluated the performance of the most known thresholding methods
and found that the sigmoidal thresholding was the one which achieved best results. In [18], the
authors proposed a nonthresholding method, after analyzing the drawbacks of the thresholding
methods. The method proposed in [1], although evaluated to different sources of colored noise,
the authors did not use the white Gaussian noise, furthermore, and did not perform comparisons
with thresholding methods. In order to fill these gaps and check the tendency for FFT/Wavelet
methods, the analysis follows below.

The methods were tested for four signals corrupted by white Gaussian noise, divided
into male and female voice. Among them, one signal into female voice is in English and the rest
of them are in Portuguese.

All signals used in the experiments are suggested by Test Signals for
Telecommunication Systems (ITU-T). The wavelet function used was the Daubechies function
of order 10 and the analysis was performed based on the degree of improvement of the SNR for
the processed signals. The signals were acquired at a 16 kHz sampling rate are in wav format.

The global signal-to-noise ratio (SNR) can be performed according to equation (2).
Taking the ratio between a voice and a silence segment of each signal as follows [5]:

557
SNR = 10 \cdot \log_{10} \left( \frac{\sum_{i=1}^{N-1} x_i^2}{\sum_{i=1}^{N-1} w_i^2} \right) \quad (2)

where $x_i$ is the voice segment, $w_i$ is the silence segment, and both with same length $N$.

Table 1 shows the values of SNR for the clean signals and Table 2 shows the enhancement for the same signals, after being corrupted by white noise, generating signals with 5 dB and 10 dB SNR, and processed by the cited methods.

<table>
<thead>
<tr>
<th>Clean signals</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male 1</td>
<td>42.19</td>
</tr>
<tr>
<td>Male 2</td>
<td>34.22</td>
</tr>
<tr>
<td>Female 1</td>
<td>30.90</td>
</tr>
<tr>
<td>Female 2</td>
<td>37.20</td>
</tr>
</tbody>
</table>

Table 1: SNR’s of clean signals.

<table>
<thead>
<tr>
<th>SNR (db)</th>
<th>Method</th>
<th>Male 1 5 dB</th>
<th>Male 1 10 dB</th>
<th>Male 2 5 dB</th>
<th>Male 2 10 dB</th>
<th>Female 1 5 dB</th>
<th>Female 1 10 dB</th>
<th>Female 2 5 dB</th>
<th>Female 2 10 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nonthresholding B</td>
<td>24.58</td>
<td>32.44</td>
<td>21.94</td>
<td>35.59</td>
<td>25.17</td>
<td>36.88</td>
<td>20.04</td>
<td>32.74</td>
</tr>
<tr>
<td></td>
<td>FFT/Wavelet</td>
<td>25.52</td>
<td>34.04</td>
<td>15.24</td>
<td>25.25</td>
<td>18.94</td>
<td>28.00</td>
<td>18.74</td>
<td>29.43</td>
</tr>
</tbody>
</table>

Table 2: SNR of speech enhancement by the considered methods from four speakers with 5 dB and 10 dB SNRs.

(* ) English.

For illustrative purposes, in Figure 2 it is possible to compare the SNR values of the enhanced signals as a function of the input signals. Each point associates the SNR of the noisy signal to its SNR after being processed by the indicated method. Observing Figure 2, it can be noted that the thresholding method had the worst performance for the four signals, while nonthresholding B provided better results for almost all signals.

Figure 2: SNR of enhanced speech as a function of the corrupted speech.

Legend: M 1 (Male 1), M 2 (Male 2), F 1 (Female 1) e F 2 (Female 2).
According to [5, 8, 18], the SNR values of the enhanced signals should be close to the clean ones. Great differences among the SNRs of the clean and the processed signal means that the denoising methods distorted the voice segments of the signal or did not remove noise as sufficiently desired.

5. Conclusion

In this paper, a brief review in the literature of the most recent speech enhancement methods was performed. Based on the results, it could clearly be seen the evolutionary process of the methods based on wavelets. The earliest wavelet methods that emerged in the literature were based on the thresholding the signals. They still have some drawbacks and were surpassed by the methods considered nonthresholding. In addition to performing a uniform noise reduction along the signal, the nonthresholding methods generated, in average, better SNR results for the enhanced signal which indicate that they could conduct to better psychoacoustics quality. Nowadays, the methods which emerged are called FFT/Wavelet. Even though the method used in this work did not stand out on the nonthresholding methods, but, it showed very promising, being a tendency for future research.

Another important fact is that the ability of the DWT in extracting coherent structures of a signal makes it a powerful tool for the analysis of non-stationary signals. Considering this wavelet characteristic, there is a tendency of combining wavelet and other signal processing approaches to get better results on wavelet speech enhancement methods.

Acknowledgments

The authors would like to thank the financial support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References


