WAVELET PACKET BASED VOICED / UNVOICED CLASSIFICATION IN NOISY ENVIRONMENT

Zied LACHIRI and Noureddine ELLOUZE
Laboratoire LSTS, Département de Génie Electrique
Ecole Nationale d'Ingénieurs de Tunis
Campus Universitaire, BP 37, 1002, Le Belvédère, Tunis, Tunisie
Tel: +00216 71 874700; fax: +00216 71 872729
e-mail: zied.lachiri@enit.rnu.tn, N.Ellouze@enit.rnu.tn

ABSTRACT

This paper describes a new robust voiced/unvoiced classification algorithm, using an appropriate wavelet packet decomposition of the speech signal. The classification is achieved by generating a correlation model of different subbands signals derived from a tree structured filter banks. The wavelet packet tree is constructed by cascading the basic two channel perfect reconstruction filters into the desired levels. To investigate the accuracy of the proposed technique, we conduct experiments using the TIMIT speech database. We add to these speech signals real world noise at various SNR. Experimental results show the accuracy of the proposed technique especially in low SNR's ($\leq 10dB$).

1 INTRODUCTION

Speech Classification can be regarded as a procedure that allows the endpointing of segments of speech from surrounding areas of speech and non speech. It plays an important role on diverse applications dealing with speech. Moreover, correct classification is crucial to the success of speaker recognition systems and adaptive speech enhancement algorithms which typically behave completely different during speech than during noise. This is true for both single sensor systems as well as for multi-sensor adaptive algorithms.

Several established Algorithm's have been used in the detection and classification of speech, they are essentially based on waveform processing (short time energy, zero crossing rate, combination of energy and zero crossing) [6], spectral estimation [5] and correlation processing [6]. In general, the parameters used in these algorithms are based on time averages over a fixed length window. Therefore, the time resolution of these algorithms depends on the choice of the window length and can not be matched to the time characteristics of the speech signal. For example, the detection of transients need high time resolution. Whereas, during stationary and periodic frames a longer analysis window, can be more efficient to extract the important signal features. Further disadvantages are either a large computational complexity or the presence of background noise especially under low SNR circumstance. So improvement in noisy environment is still a remaining subject.

Commonly, speech sound is considered to be a signal whose component localisation vary widely in time and frequency, it contains both high/low frequency components and short/large duration sounds. Therefore it's important to decompose speech into waveforms whose time frequency properties are adapted to its local structures [7]. Considering it's mathematical property and the capability to model speech sounds, the wavelet packet [12] is well suited to this type of expansion. The wavelet packet transform is an analysis method that offers more flexibility in adapting time and frequency resolution to the input signal. This flexibility is achieved by correlating the input signal with basis functions that are scaled and shifted versions of a so called mother wavelet which itself is a bandpass function.

This papers focuses on speech classification in real word noise. Section 2, introduces a brief overview on the wavelet transform and the subband wavelet packet decomposition. In section 3, we describe a new voiced unvoiced classification algorithm in noisy environment. This technique based on time and frequency feature uses a correlation model of different subbands speech signals derived from a tree structured filter bank properly choosed to extract the speech signal characteristics. Section 4, presents the effectiveness of the proposed method and discusses the simulation results. finally, the main conclusion of our work are summarised.

2 WAVELET PACKET SUBBAND DECOMPOSITION

Wavelet transform [2] [4] [10] was recently introduced as an alternative technique for analysing non stationary signal. It provide a new way for representing signal into well-behaved expression that yields useful properties. The wavelet is a square integrable function well localised in time and frequency, from which we can extract all basis functions by using variations of the basic wavelet obtained by time shifting and scaling.

The continuous wavelet transform of signal x relative

to the basic wavelet is given by:

$$W_\psi x(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} x(t)\psi^* \left(\frac{t-b}{a}\right) \qquad (1)$$

where a, b $(a, b \in \mathbb{R}; a \neq 0)$ are respectively the translation and scale parameters. Furthermore, if the basic wavelet satisfy the admissibility condition [4], then, the wavelet reconstruction formula is:

$$x(t) = \iint_{\mathbb{R}} W_\psi x(a,b)\psi_{a,b}(t)\frac{dadb}{a^2} \qquad (2)$$

The continuous wavelet transform is essentially employed to derive properties, however, discrete forms are necessary for practical applications. Discrete time implementation of wavelet is based on a tree structure which uses a single basic building block repeatedly until the desired decomposition is accomplished. This basic unit uses techniques of multi-rate signal processing [3] and consists of a low and a high pass filter followed by a down-sampling unit. The first stage splits the signal into a high-pass and low-pass band, each of which is spread to full band by the subsequent downsampling. Given this spreading that accompanies downsampling, the second stage can be viewed as simply splitting the low-pass portion of the original signal into halves. Each stage of the discrete wavelet transform thus splits the low-pass spectrum from the previous stage; This results in an octave-band filter bank in which the sampling rate of a subband is proportional to its bandwidth.

The wavelet analysis is sometimes inefficient because it only partitions the frequency axis finely toward the low frequency. the wavelet packet transform [12] constitutes a solution that permits a finer and adjustable resolution of frequencies at high frequencies and gives a rich structure that allows adaptation to particular signals or signals classes [2]. Unlike the wavelet transform, the wavelet packet transform divides the low and the high frequency subband, resulting in tree structured filter bank called a wavelet packet filter bank. This transformation creates a division of the frequency domain to represents the signal optimally.

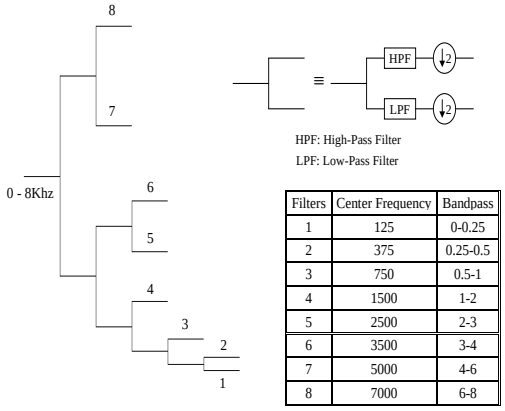
3 VOICED / UNVOICED CLASSIFICATION

3.1 Speech Subband Decomposition

Speech can simply be classified as voiced, unvoiced and silence. Voiced speech is quasi-periodic in the time domain and harmonically structured in the frequency domain while unvoiced speech is random like and broadband. The voiced sound is frequency limited signal which has most of the energy in the low frequency range, less than 1 Khz, whereas the energy of unvoiced speech is usually concentrated at the high end of frequency scale ($\geq$ 3Khz) [11]. If we want to get a discrimination of the voiced and unvoiced sounds we must derive benefit from the information contained in those bands where the voiced sound or the unvoiced sound is dominant compared with the other sounds.

It is known that most of the speech signal power is contained around the first formant. the statistical results for many vowels of adult males and females indicates that the first formant frequency doesn't exceed 1 Khz and doesn't below 100Hz approximately. In addition, pitch frequency lies in normal speech between 80 and 500 Hz.
Based on these spectral behaviours, we suggest to decompose the speech signal $x(t)$ into 8 subband wavelet packet tree:

$$\begin{aligned} x(t) &= \sum_{i=1}^{8} W_\psi x(i)\psi_i(t) \\ W_\psi x(i) &: \text{Wavelet packet (WP) transform of x} \\ i &: \text{subband frequency index}(i = 1, 2, ..., 8) \\ \psi_i &: \text{WP function of the ith subband} \end{aligned}$$



Filters	Center Frequency	Bandpass
1	125	0-0.25
2	375	0.25-0.5
3	750	0.5-1
4	1500	1-2
5	2500	2-3
6	3500	3-4
7	5000	4-6
8	7000	6-8

Figure 1: *8 subband wavelet packet tree covering* $0 - 8Khz$ *and their parameters: Center frequency (Hz) and Bandpass (Khz)*

The proposed tree assigns more subband in low frequency which normally contain large portions of the signal energy. The wavelet packet transform is computed for the given wavelet tree, which result in a sequence of subband signals or equivalently the wavelet packet transform coefficients, at the leaves of the tree. In effect, each of these subband signals contains only restricted frequency information due to inherent band-pass filtering. The filter bank that implements the wavelet packet decomposition and the time frequency tilling are given respectively in Figure 1 and 2 (The depicted decomposition scheme is for a sampling rate $f_e = 16Khz$).

3.2 Subband Crosscorrelation

The speech signal is highly correlated in case of voiced speech. This fact make it possible to track the uncor-

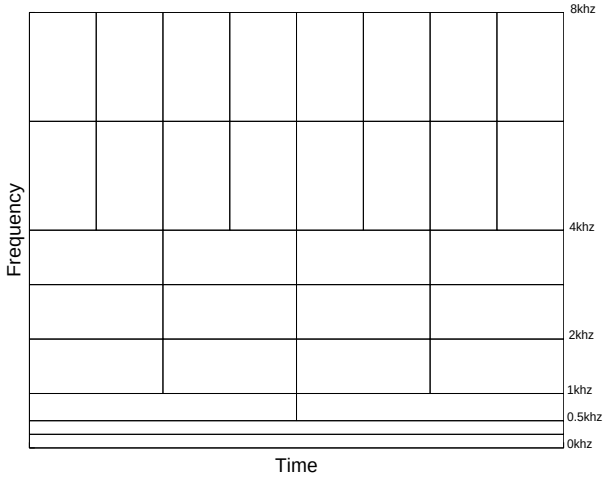


Figure 2: *time frequency tiling of the proposed wavelet packet tree*

related portions and extract the pure speech segments. This procedure is still effective to detect the voice activity in speech signal both in noise and noise free. In effect, any transition between a silence and voiced sound or unvoiced sound can be identified by the Subband Crosscorrelation Analysis (SUB-CRA) [9] between different subband signals obtained via wavelet packet subband decomposition. This technique give the maximum reliable correlation representation between the subband signals and get the highest immunisation to noise. Moreover, the nature of the wavelet packet decomposition makes it possible to control the signal into many bands each has a portion of the noise power, which is much less than the total noise power distributed in all bands especially in the case of normal distribution of noise.

The algorithm begins by splitting the speech signal $x(n)$ into windows $x_w(n) = x(n-m)w(m)$. Each window is passed through an appropriated filterbank to extract the wavelet packets parameters. The Subband Crosscorrelation Analysis (SUB-CRA) is performed using different filters responses (figure 1): filters 1, 2 and 3 are selected to detect the voiced segments and filters 6, 7 and 8 are selected to locate the unvoiced segments. The selection of the frequency bands is based on the speech behaviour which indicates that the most power of the voiced sound and the unvoiced sound reside respectively in the low frequency (≤ 1 Khz) and the high frequency bands (≥ 3 Khz).

After selecting the filters responses, the crosscorrelation functions R^k_{1-2}, R^k_{2-3}, R^k_{1-3}, R^k_{6-7} and R^k_{7-8} between the filters outputs, are generated for each frame k, where:

$$R^k_{i-j}(l) = \sum_{l_1=0}^{2N-1} x^i_w(l_1) x^j_w(l_1 + l). \qquad (3)$$

and k, N define respectively the frame rank and the length of the subband signal x^i_w.

To generate any crosscorrelation function defined above, a simple interpolation technique is used to insert points between the wavelet packets parameters to expand them in each frequency band to the window length. The frames of the all crosscorrelation parameters are concatenated, then the absolute value of the points is taken and smoothed using moving average of N points length. Consequently, we obtain 5 envelope function R_{1-2}, R_{2-3}, R_{1-3}, R_{6-7} and R_{7-8}, where:

$$R_{i-j}(l) = \sum_{k=1}^{L+1} R^k_{i-j}\left(l - (k-1)N\right). \qquad (4)$$

L: total number of frames.

We choose two envelope function R_1 and R_2. R_1 is selected as the maximum energy contribution from R_{1-2}, R_{2-3}, R_{1-3} and R_2 is selected, too, as the maximum energy contribution from R_{6-7}, R_{7-8}.

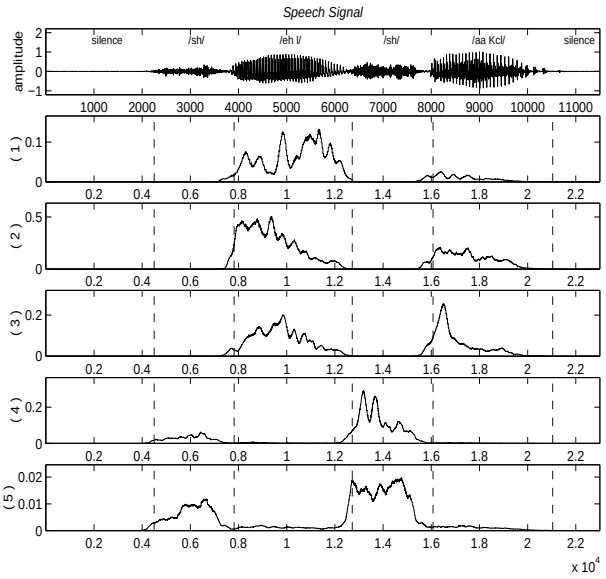


Figure 3: *The smoothed crosscorrelation functions (1) R_{1-2}, (2) R_{2-3}, (3) R_{1-3}, (4) R_{6-7} and (5) R_{7-8} of the speech signal depicted in the first subfigure*

As shown in figure 3, the energy changes can easily be detected. correlating the energy contents of the same signal in two different frequency levels generates the curves shown.

3.3 Experimental Results

In order to evaluate the performance of classifying the speech sounds by crosscorrelation method, experiments by computer simulation is carried out. The Speech signals uttered by male and female speakers are obtained from TIMIT corpus. The test set consists of a total of 465 frames of data sampled at $16Khz$ rate. 196 and 144 frames are manually labelled as voiced and unvoiced segment. Experiments were conducted by adding real world

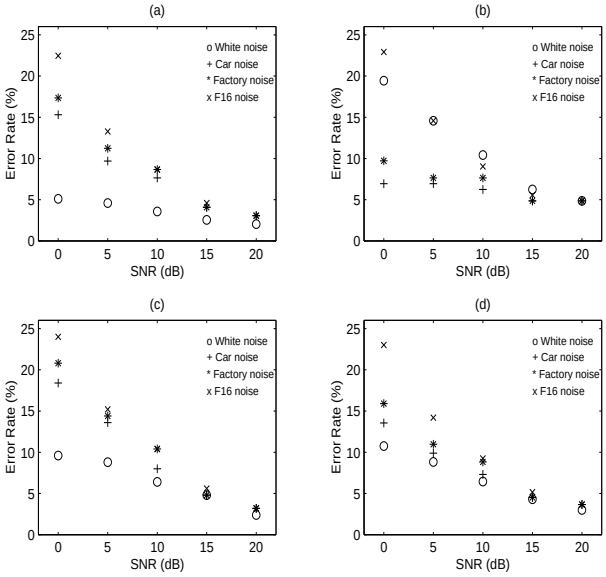


Figure 4: Performance of Subband Crosscorrelation Analysis. (a) error rate in detecting voiced sounds, (b) error rate in detecting unvoiced sounds, (c) error rate in detecting silence and (d) global error rate.

noise: white noise, factory noise, volvo noise and F16 jet engine noise, with different Signal to Noise ratio (20db, 10db, 5db and 0db. The other details in the experiments are as follows: window size is 32ms, the window shift is 16ms and the mother wavelet is Daubechies 10.
The discrimination of the speech segments (voiced sound and unvoiced sound) from noise is conducted using a comparison with an appropriate threshold, which is generated exploiting the first frames of the correlation model.

The results for noisy speech are plotted in figure 4, which contains the detection error respectively for the voiced segments , the unvoiced segments and silence. The analysis of figure 4 shows that an error rate less than 6% (white noise) is achieved for voiced sound detection in white noise even in the hard cases ($SNR = 0dB$). Whereas, in detecting both the unvoiced sound and silence, we observe in the same noisy environment, an error rate greater than 9%.
In the case where the speech signal is corrupted by factory noise or volvo noise, the opposite phenomena is observed. This noting indicates that the performances of the proposed technique for detecting voiced sounds are more sensitive to narrow band noise. We note also that for the all SNR's, the technique generate more detection error where speech signal is contaminated by F16 jet engine noise which exhibits significant non stationnarity in power and frequency content.

4 CONCLUSION

We propose a robust voiced/unvoiced classification algorithm in noisy environment, using an appropriate wavelet packet decomposition of the speech signal. Classification is achieved by subband crosscorrelation analysis generated using a correlation of different subbands signals derived from a tree structured filter banks. Based on experiment results it is shown that the proposed method can detect accurately the voiced and unvoiced sounds, even in low SNR ($< 10db$).

5 REFERENCES

[1] N. Abdel Kader and A. M. Refat. "Voiced / Unvoiced Classification using Wavelet based Algorithm", ICSPAT, 1998.

[2] C. S. Burrus, R. A. Gopinath and H. Guo. "Introduction to Wavelets and Wavelet Transforms: A Primer", Prentice Hall, 1998.

[3] R. Crochiere and L. R. Rabiner. Multirate Digital Signal Processing, Prentice Hall, Englewood Cliffs, NJ, 1983.

[4] I. Daubechies, Ten lectures on wavelets, SIAM Press, 1992.

[5] J. Haigh and J. S. Mason. "A Voice Activity Detector based on Cepstral Analysis", Proc. Eurospeech-93, Berlin, Germany, 1993.

[6] W. J. Hess. Pitch and Voicing Determination, in Advances in Speech Signal Processing, ed. S. Furui and M.M. Sondhi, Marcel Dekker, 1992.

[7] Z. Lachiri and N. Ellouze. "Speech Representation Based on Wavelet Function" *CESA 98 IMACS Multiconference, Computational Engineering in Systems Applications*, pp:192-195, Nabeul-Hammamet, Tunisia, April 1998.

[8] Z. Lachiri. "Voiced / Unvoiced Classification based on Wavelet Packet Transform" *Tunisian German Conference on Smart Systems and Devices (SSD)*, pp: 660-665, 27-30 Mars 2001, Hammamet, Tunisia.

[9] Z. Lachiri and N. Ellouze. "Voiced / Unvoiced Classification using Subband crosscorrelation Analysis" *International Congress of Acoustics, ICA2001*, 2-7 September 2001, Rome, ITALY.

[10] S. Mallat. A Wavelet Tour of Signal Processing, Second Edition, Academic Press, 1999.

[11] L. R. Rabiner and R. Schafer. Digital Processing of Speech Signals, Prentice Hall, Englewood Cliffs, NJ, 1978.

[12] M. V. Wickerhauser. Adapted Wavelet Analysis from Theory to Software, A. K. Peters, Wellesley, Massachusetts, 1994.