Subband Crosscorrelation Analysis based Speech Classification in Noisy Environment

ZIED LACHIRI [1] and NOUREDDINE ELLOUZE [2]
[1] Département de Physique et Instrumentation
Institut National des Sciences Appliquées et de Technologies
BP 676, 1080, Centre Urbain Cedex, Tunis
TUNISIA
[2] Département de Génie Electrique
Ecole Nationale d'ingénieurs de Tunis
BP 37, 1002, le Belvédère, Tunis
TUNISIA
zied.lachiri@en .rn tn, N.Ellouze@ .rn tn

Abstract: - This paper presents a new algorithm for robust speech classification in adverse conditions, using an appropriate wavelet packet decomposition of the speech signal. The classification is achieved by generating a Sub-band Crosscorrelation Analysis of different subbands signals derived from a tree structured filter banks. The performance of the proposed technique is evaluated on speech signal with real world noise, added to it, at various SNR. Experimental results show the accuracy of the proposed technique especially in low signal to noise circumstance $< 10\ dB$.

Key-Words: - Wavelet packet expansion, Crosscorrelation function, Voiced / unvoiced / Silence Classification.

1 Introduction

Speech Classification can be regarded as a procedure that allows the end-pointing of segments of speech from surrounding areas of speech and non speech. It plays an important role on diverse applications dealing with speech. Moreover, accurate speech classification is required for the success of speaker recognition algorithms and many speech enhancement systems.

Several established Algorithm's have been used in the detection and classification of speech, they are essentially based on waveform processing (energy, zero crossing rate) [5], correlation processing [5] and spectral estimation [4]. 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 needs high time resolution. Whereas, during stationary and periodic frames a longer analysis window, can be more efficient to extract the important signal features. Further disadvantage is 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 localization varies 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. Considering its mathematical property and the capability to model speech sounds, the wavelet packet [10] 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 band pass function.

This paper focuses on speech classification in noisy environment. In section 2, we introduce 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 choozed to extract the speech signal characteristics. In section 4, we present the effectiveness of the proposed method and discuss the simulation results. Finally, the main conclusions of our work are summarized in section 5.

2 Speech Decomposition via Wavelet Packet

Wavelet transform [8] was recently introduced as an alternative technique for analyzing non stationary signal.

It provides a new way for representing signal into well-behaved expression that yields useful properties.
The continuous wavelet transform of signal $x(t)$ 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 $(a,b \in IR,\ a \neq 0)$ are respectively the translation and scale parameters.
This 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 [2] and consists of a low and a high pass filter followed by a down-sampling unit. 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 [10] 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. 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.

2.1 Speech 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 1Khz, whereas the energy of unvoiced speech is usually concentrated at the high end of frequency scale ($\geq 3Khz$) [9]. 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 1Khz and doesn't below 100Hz approximately. In addition, pitch frequency lies in normal speech between 80 and 500Hz. Based on these spectral behaviors, we suggest to decompose the speech signal $x[n]$ into 8 subband wavelet packet tree:

$$x[n] = \sum_{i=11}^{8} W_\psi x(i)\psi_i[n] \qquad (2)$$

where $i = 1, 2, \ldots\ldots, 8$, $W_\psi x(i)$ and $\psi(i)$ represent respectively the subband frequency index, the Wavelet Packet (WP) Coefficients and the Wavelet Packet function of the i'th subband.

Filter	Center Frequency (Hz)	Band-pass (Hz)
1	125	0 – 250
2	375	250 – 500
3	750	500 – 1000
4	1500	1000 – 2000
5	2500	2000 – 3000
6	3500	3000 – 4000
7	5000	4000 – 6000
8	7000	6000 – 8000

Tab.1 8 subband wavelet packet tree covering 0 - 8Khz and their parameters: Center frequency Bandpass

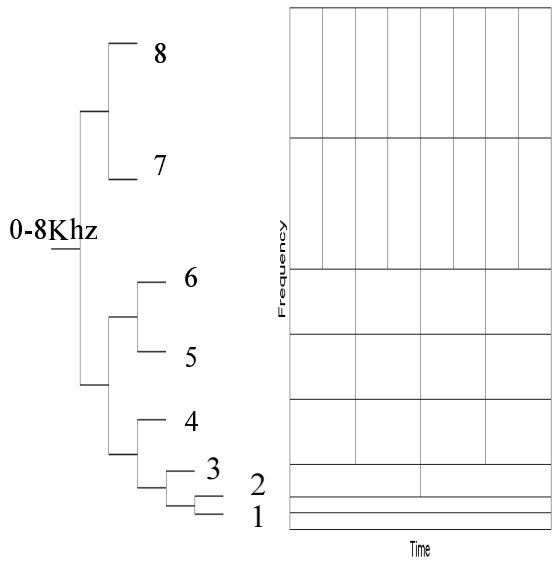

Fig.1 Time frequency tiling of the proposed wavelet packet tree.

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 (a) and (b) (The depicted decomposition scheme is for a sampling rate $f_e = 16Khz$).

3 Subband Crosscorrelation Analysis

The speech signal is highly correlated in case of voiced speech. This fact makes it possible to track the uncorrelated 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 [6] between different subband signals obtained via wavelet packet subband decomposition. This technique gives the maximum reliable correlation representation between the subband signals and gets the highest immunization 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 filter bank to extract the wavelet packets parameters. The Subband Crosscorrelation Analysis 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 behavior which indicates that the most power of the voiced sound and the unvoiced sound reside respectively in the low frequency ($\leq 1Khz$) and the high frequency bands ($\geq 3Khz$).

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

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

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

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 low pass filter. Consequently, we obtain 5 envelopes functions $R_{1-2}[n]$, $R_{2-3}[n]$, $R_{1-3}[n]$, $R_{6-7}[n]$ and $R_{7-8}[n]$, where:

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

L: total number of frames.

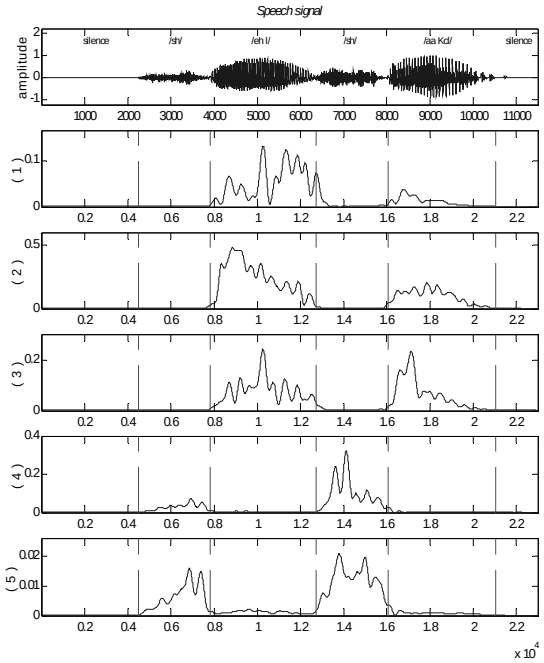

Fig.2 The smoothed crosscorrelation functions (1) $R_{1-2}[n]$, (2) $R_{2-3}[n]$, (3) $R_{1-3}[n]$, (4) $R_{6-7}[n]$ and (5) $R_{7-8}[n]$ of the speech signal depicted in the first subfigure.

The vocal signal described by figure 2, is constituted by two voiced segments and two unvoiced segments, from where the existence of five zones of transitions. The application of the wavelet packet transform according to the proposed tree [6] permits to get the decomposition coefficients. As shown in figure 2, the energy changes can easily be detected. Correlating the energy contents of

the same signal in two different frequency levels generates the curves shown.

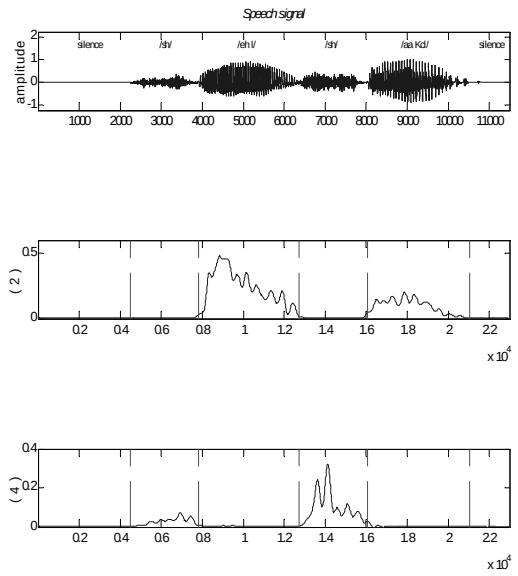

Fig.3 The choosed crosscorrelation functions (1) $R_1[n]$ and (2) $R_2[n]$ of the speech signal depicted in the first subfigure

After applying subband crosscorrelation analysis, we choose two envelope function $R_1[n]$ and $R_2[n]$. $R_1[n]$ is selected as the maximum energy contribution from $R_{1-2}[n]$, $R_{2-3}[n]$, $R_{1-3}[n]$ and $R_2[n]$ is selected, too, as the maximum energy contribution from $R_{6-7}[n]$, $R_{7-8}[n]$. 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 functions $R_1[n]$ and $R_2[n]$.

The research of every class (voiced, unvoiced and silence) is completely carried out in independent manner. Two types of location conflicts can appear: 1) Obtaining disconnected segments and 2) Obtaining overlapped segments.

To solve these conflicts and produce a coarse segmentation of the signal it's necessary to take into account the locally results location of each stage. The proposed solution is inspired from the works of Fohr [3] and Haigh [4]:

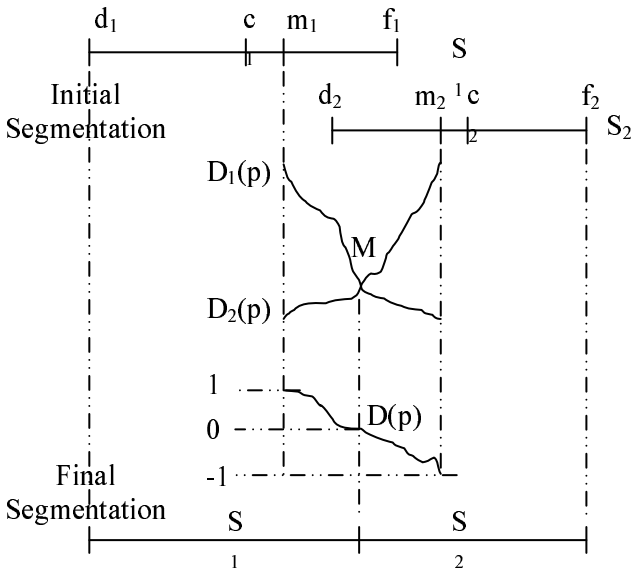

Fig.4 Case of overlapped segments. The distances $D_1(p)$, $D_2(p)$ and $D(p)$.

Case of disconnected segments: The portions of the signal that are not classified in one of the two explicitly classes (sought-after) are cataloged "another" and are supposed to correspond to "intervals of silence". This technique solves the disconnected segment problem and all portions of the signal are thus cataloged by at least one class. The remaining contradictory cases are cases of overlap segments.

Case of overlapped segments: The algorithm tries to get some limit between segments. The figure 4 shows a typical case of overlapped segments $S_1(d_1, m_1, f_1)$ and $S_2(d_2, m_2, f_2)$ where d_i, f_i and m_i mark respectively the beginning sample of the segment i, the ending sample of the segment i and the maximum between the center c_i and the border of the segment i that presents a conflict (either d_i or f_i).

In order to solve the conflict of overlap, we propose to determine a border between segments S_1 and S_2 in the interval $[m_1, m_2]$. For each sample p brings in m_1 and m_2 the distances (absolute middle value) $D_1(p, m_1) = |R_1(p) - R_1(m_1)|$ and $D_2(p, m_2) = |R_2(p) - R_2(m_2)|$, are calculated respectively with the samples m_1 and m_2, where $R_1(k)$ and $R_2(k)$ define the chosen crosscorrelation functions for located the voiced zone and unvoiced zones. The distances $D_1(p, m_1)$ and $D_2(p, m_2)$ permit to determine the function thereafter:

$$D(p) = \frac{D_1(p, m_1) - D_2(p, m_2)}{D_1(p, m_1) + D_2(p, m_2)} \qquad (5)$$

The evolution of this function forms a curve of which an example is represented in the figure 4. The curve $D(p)$ passes by a point M for which the distance is $D(M) = 0$ and the temporal abscissa is t_M. This point constitutes the border of segmentation between segments S_1 and S_2. Thus, one gets a sequence of two adjacent segments.

4 Results and Discussion

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 2156 frames of data sampled at 16 Khz rate. 845 and 683 frames are manually labeled as voiced and unvoiced segment. Experiments were conducted by adding real world noise: Bursting noise, Factory noise, Volvo noise and F16 jet engine noise, with different Signal to Noise Ratio (15dB, 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.

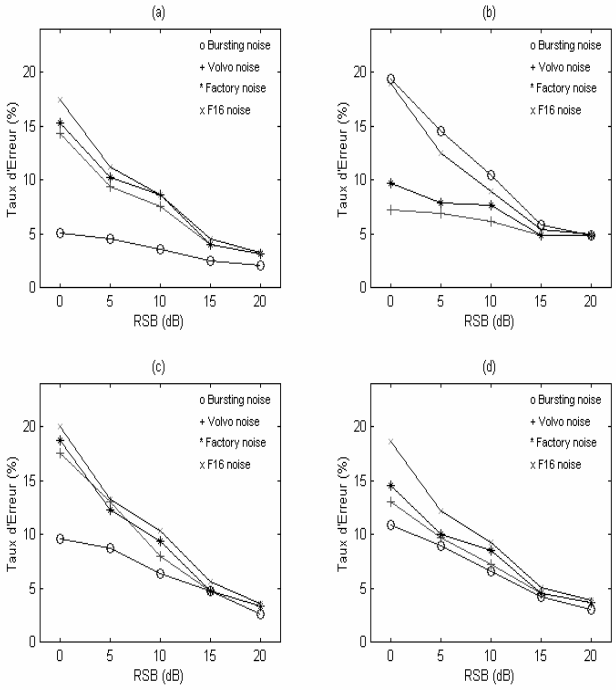

Fig.5 Performance of the proposed method. (a) error rate in detecting voiced sound, (b) error rate in detecting unvoiced sounds, (c) error rate in detecting silence and (d) global error rate.

The results for noisy speech are depicted in figure 5, which contains the detection error respectively for the voiced segments, the unvoiced segments and silence. Errors in the voiced and unvoiced columns indicate missed detections, whereas errors in the silence column indicate false alarms. The analysis of the figure 5-(a) shows that an error rate less than 6% (Bursting noise) is achieved for voiced sound detection in Bursting noise even in the hard cases (SNR=0dB). Whereas, in detecting both the unvoiced sound and silence (figure5 (b) , (c)), we observe in the same noisy environment and the same condition (0dB), an error rate detection of unvoiced sound and silence less respectively than 20% and 10%.

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 stationary 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).

References:

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

[2] R. Crochiere and L. R. Rabiner, *Multi-rate Digital Signal Processing*, Prentice Hall, 1983.

[3] D. Fohr, *APHODEX, un système expert en décodage acoustico phonétique de la parole continue*, Thèse, Université Nancy 1, CRIN, Mars 1986.

[4] A. J. Haigh, *Voice Activity Detection for Conversational Analysis*, Master Thesis, College of Swansea, Wales, 1994.

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

[6] Z. Lachiri and N. Ellouze, Voiced / Unvoiced Classification using Subband crosscorrelation

Analysis, International Congress of Acoustics, ICA2001, September 2001, Rome, ITALY.
[7] Z. Lachiri and N. Ellouze, *Wavelet packet based Voiced / Unvoiced Classification in noisy environment*, EUSIPCO2002, September 2002, Toulouse, France.
[8] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1999.
[9] T. F. Quatieri, *Discrete Time Speech Signal Processing*, Prentice Hall, 2002.
[10] M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, A. K. Peters, Wellesley, Massachusetts, 1994.