Open Access

Two-Feature Voiced/Unvoiced Classifier Using Wavelet Transform

A.E. Mahdi* and E. Jafer

Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland

Abstract: This paper proposes a new wavelet-based algorithm for voice/unvoiced classification of speech segments. The classification process is based on: 1) statistical analysis of the energy-frequency distribution of the speech signal using wavelet transform, and 2) estimation of the short-time zero-crossing rate of the signal. First, the ratio of the average energy in the low-frequency wavelet subbands to that of highest-frequency wavelet subband is computed for each time segment of the pre-emphasised speech using a 4-level dyadic wavelet transform, and compared to a pre-determined threshold. This is followed by measuring the zero-crossing rate of the segment and comparing it to a threshold determined by a continually up-dated value of the median of the zero-crossing rates of the speech signal. An experimentally verified criterion based on the results of the above two comparison processes is then applied to obtain the classification decision. The performance of the algorithm has been evaluated on speech data taken from the TIMIT database, and is shown to yield high classification accuracy and robustness to additive noise.

INTRODUCTION

Classification of speech into voiced and unvoiced (V/UV) regions is an important and necessary process in many high-quality speech processing applications, such as speaker identification and verification, speech coding and synthesis, and speech enhancement using wavelet thresholding. Essentially, the classification involves the determination of whether the speech is produced by vibration of the vocal cords [1]. The process can effectively be performed using a single feature or parameter which is closely associated with the voicing and non-voicing activities of the speech. However, due to the fact that the range of values of any single speech parameter overlaps between different regions, the accuracy of a single-feature V/UV classification method would be limited [2]. In the context of a single-feature approach, the problem of speech classification has been traditionally associated with the determination of pitch frequency. However, the vibration of the vocal cords does not necessarily result in periodicity of the speech signal and, therefore, a failure in detection of periodicity in some voiced regions would result in V/UV classification errors [4]. Several classification algorithms that employ more than one speech feature, such as the pitch frequency, zero-crossing rate, Cepstral peaks and/or short-time energy distribution, have been reported to offer better accuracy compared to single-feature approaches [1-3].

In recent years, the wavelet transform (WT) has been shown to provide a powerful and computationally- efficient tool for a variety of signal and speech processing applications [5-9]. These include a number of single-feature V/UV classification techniques based on a pitch determination approach [10, 11], or based on the per wavelet subband average energy distribution [12-14]. In this paper, we propose a new technique for V/UV speech classification using two features of the speech signal: (a) the frequency distribution of the average energy and (b) zero-crossing rate, for each speech segment. The performance of the proposed technique has been evaluated using a large database of both clean speech and speech degraded with additive noise.

Following this introduction, the second Section gives an overview of the discrete wavelet transform (DWT). The detailed description of the proposed V/UV classification algorithm is given next. In the forth Section, the various experimental results of the performance evaluation of the algorithm are illustrated and discussed. The article concludes by giving a summary of the proposed algorithm and findings of the evaluation process.

WAVELET TRANSFORM

The wavelet transform is a fundamental mathematical tool for multi-resolution or multi-scale decomposition of signals, capable of providing the time and frequency information simultaneously. It involves mapping signals onto time-scale space with superior time-frequency localisation as compared to the short-time Fourier transform (STFT), which in some cases cannot show the locations of discontinuity in signals. The continuous wavelet transform (CWT) is based on the correlation of a given continuous-time signal with a set of filters defined by a family of basis wavelet functions. The two-dimensional parameterisation is achieved by generating a set of basis wavelets by dilation and translation of a single function $\psi(t)$, known as the 'mother wavelet', using the following relation:

$$\psi_{a,b}(t) = |a|^{\frac{1}{2}} \psi(\frac{t-b}{a}) \qquad (1)$$

where a and b are real and continuous parameters known as the scale or dilation parameter and the shift or translation parameter, respectively, and $\psi_{a,\,b}(t)$ is the set of wavelet basis functions referred to as a wavelet family [15]. Since the pa-

*Address correspondence to this author at the Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland; Tel: (+)353-61-213492; Fax: (+)353-61-338176; E-mail: Hussain.Mahdi@ul.ie rameters *a* and *b* are continuous valued, the transform is called continuous wavelet transform. The definition of classical wavelets as dilates of one function means that high frequency wavelets correspond to $a < 1$, or narrow width, while low frequency wavelets have $a > 1$, or wider width. By applying wavelet transform, a function *f(t)* is expressed as a linear combination of scaling and wavelet functions. Both the scaling functions and the wavelet functions are complete sets. However, it is a common practice to employ both wavelet and scaling functions in the transform representation of the function. In general, the scale and shift parameters of the wavelet family are discretised by sampling on a gird, such that $a = a_0^{\,j}$ and $b = kb_0 a_0^{\,j}$, where *j* and *k* are integers. Hence, the wavelet function family with discretised parameters is expressed as:

$$\psi_{j,\,k}(t) = a_0^{-j/2}\,\psi(a_0^{-j}t - k\,b_0) \tag{2}$$

where $\psi_{j,\,k}(t)$ is referred to as the basis functions of the discrete wavelet transform (DWT), a_0 and b_0 are the sampling intervals for *a* and *b*, and $j,\,k \in \boldsymbol{Z}$ where $\boldsymbol{Z}$ is the set of all integers. However, although the transform is called DWT, the time variable of the transform is still continuous. The DWT coefficients of a continuous time function, *f(t)*, are thus defined as:

$$d_{j,k} = \; < f(t), \psi_{j,k}(t) > = a_0^{-j/2} \int f(t)\, \psi(a_0^{-j}t - kb_0)\, dt \tag{3}$$

When the DWT set $\psi_{j,k}(t)$ is complete, *f(t)* is expressed in terms of the wavelet representation as:

$$f(t) = \sum_j \sum_k < f(t), \psi_{j,k}(t) > \; \psi_{j,k}(t) \tag{4}$$

Accordingly, we may express any continuous-time function, *f(t)*, as a complete series with *L* finite resolutions (finite number of decomposition stages) of wavelets and scaling functions using a dyadic wavelet grid, whereby $a_0 = 2$ and $b_0 = 1$, as:

$$\begin{aligned} f(t) = &\sum_{k=-\infty}^{\infty} c_{L,k}\, 2^{-L/2}\; \phi_{L,k}(2^{1/L} - k) + \\ &\sum_{j=1}^{L} \sum_{k=-\infty}^{\infty} d_{j,k}\, 2^{-j/2} \psi(2^{t/j} - k) \end{aligned} \tag{5}$$

where the scaling coefficients are similarly defined as:

$$c_{L,k} = < f(t), \phi_{L,k}(t) > = \int f(t)\; 2^{-L/2} \phi(\frac{t}{2^L} - k)\, dt \tag{6}$$

where

$\phi_{L,k}(t) = 2^{-L/2} \phi(2^{-L}t - k)$, $\psi = 2\sum_k h_1(k)\,\phi(2t - k)$ and $\phi = 2\sum_k h_0(k)\,\phi(2t - k)$, with $h_0(k)$ and $h_1(k)$ being the scaling and wavelet filters, respectively [12, 16]. The above parameterisation of the time or space location by *k* and the frequency or scale by *j* has been proven to be very effective in the analysis of non-stationary signals [16].

VOICED/UNVOICED CLASSIFICATION ALGORITHM

The proposed voiced/unvoiced (V/UV) is based on the computation of two features of the speech signal: (a) the average per subband energy distribution in the wavelet domain and (b) the zero-crossing rate. A block diagram representation of the algorithm is given in Fig. (**1**). First the pre-emphasised speech signal, sampled at 8 kHz (a commonly used sampling rate for telecommunication applications), is segmented into 20 ms long segments with 50% overlap using a sliding Hamming window. Prior to that, an optional noise suppression stage may be used to de-emphasis the out-of-band noise in the cases where the signal-to-noise ratio (SNR) of the speech is below 5 dB. This is achieved *via* the application of a Butterworth band-pass filter with lower and upper cut-off frequencies of 200 Hz and 3400 Hz, respectively. After the pre-processing stage, the following analysis and feature-extraction processes are implemented:

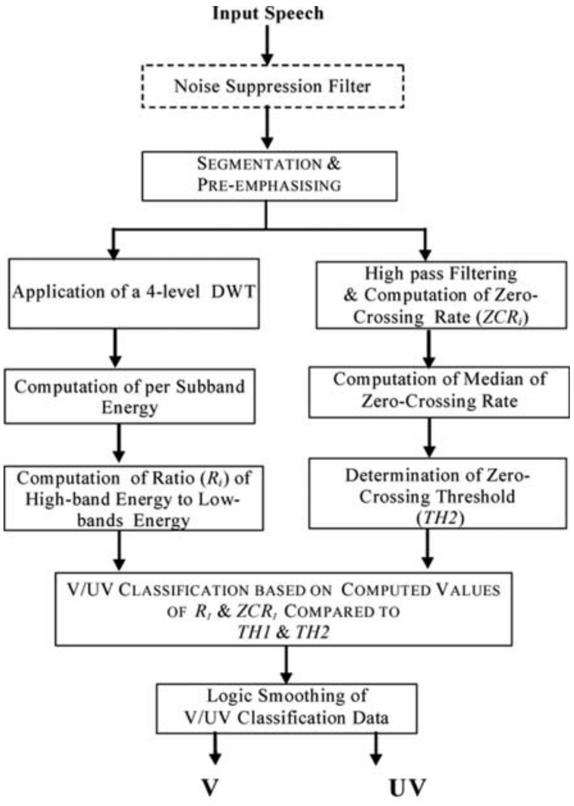

Input Speech
Noise Suppression Filter
SEGMENTATION & PRE-EMPHASISING
Application of a 4-level DWT
High pass Filtering & Computation of Zero-Crossing Rate (ZCR_i)
Computation of per Subband Energy
Computation of Median of Zero-Crossing Rate
Computation of Ratio (R_i) of High-band Energy to Low-bands Energy
Determination of Zero-Crossing Threshold (*TH2*)
V/UV CLASSIFICATION BASED ON COMPUTED VALUES OF R_i & ZCR_i COMPARED TO *TH1* & *TH2*
Logic Smoothing of V/UV Classification Data
V
UV

Fig. (1). Block diagram of the proposed V/UV speech classification algorithm.

Wavelet-Based Frequency Distribution of the Average Energy

In this stage, each speech segment is decomposed into five wavelet subbands using a 4-level dyadic DWT, and the average energy of each subband in the wavelet domain is computed. In general, an unvoiced speech segment should show energy concentration in the high frequency subbands, while a voiced segment should show energy concentration in low frequency subbands of the wavelet domain. Fig. (**2**), for example, shows the result of the analysis of the energy distribution in the wavelet domain of a clean speech signal incorporating both voiced and unvoiced regions. It is clear from Fig. (**2b**), where the per segment ratio (R_i) of the accumulated average energy of wavelet low-subbands to that of the wavelet highest-subband in the *i*th segment is shown, that a reliable V/UV separation could be achieved by comparing the average energy of the wavelet low-subbands to those of the wavelet high-subbands. Based on this and adapting an approach reported in [11], the ratio between the energy of the low subbands (i.e. below 2 kHz) to that of the highest subband (i.e. above 2 kHz) is computed and used in our algorithm as the first fundamental parameter in formulating the V/UV decision.

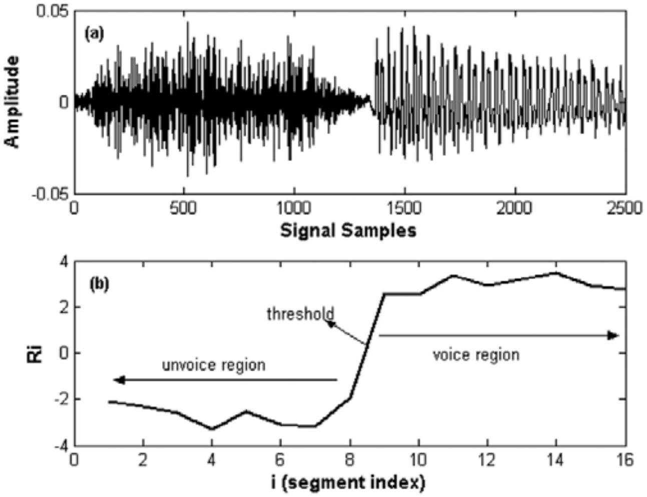


Fig. (2). (**a**) Waveform of a clean unvoiced-voiced speech signal, and (**b**) corresponding wavelet-based frequency distribution of the average energy in each segment.

Let E_i be the total energy in the *i*th speech segment and $e_{i,j}$ be the energy in wavelet subband *j* of segment *i*, such that:

$$E_i = \sum_{j=1}^{5} e_{i,j} \tag{7}$$

where j =1 represents the highest frequency subband in the wavelet domain (i.e. above 2kHz), and j = 2 to 5 represent the frequency subbands 0-2 kHz. The above-mentioned per segment energy ratio is then computed as

$$R_i = \frac{\sum_{j=2}^{5} e_{i,j}}{e_{i,1}} \tag{8}$$

The values of the parameter R_i have been examined using a large database of clean and degraded speech signals of various lengths. Accordingly, a value of 0.95 was found to be the most appropriate criterion to discriminate between voiced and unvoiced segments. Hence, the first threshold for making the required V/UV classification using the DWT energy distribution, is deemed to be

$$TH1 = 0.95 \tag{9}$$

such that if $R_i < 0.95$, the segment is most likely to be unvoiced; otherwise, the speech segment is likely to be voiced.

Zero-Crossing Rate

For discrete-time signals, a zero-crossing occurrence is detected when two successive samples have different algebraic signs. Hence, a sufficiently accurate zero-crossing count can be affected by comparing the signs of successive samples. However, in speech signals, the existence of noise, DC offset and 50 Hz hum significantly reduces the accuracy of such a simple measurement approach. In this algorithm, the speech signal is first filtered using a Chebyshev high-pass filter with a cut-off frequency of 100 Hz to minimise the above-mentioned effects. The zero-crossing rate corresponding to the *i*th segment of the filtered speech is then computed as follow:

$$ZCR_i = \sum_{n=0}^{N-1} \left| \operatorname{sgn}[x_i(n)] - \operatorname{sgn}[x_i(n-1)] \right| \tag{10}$$

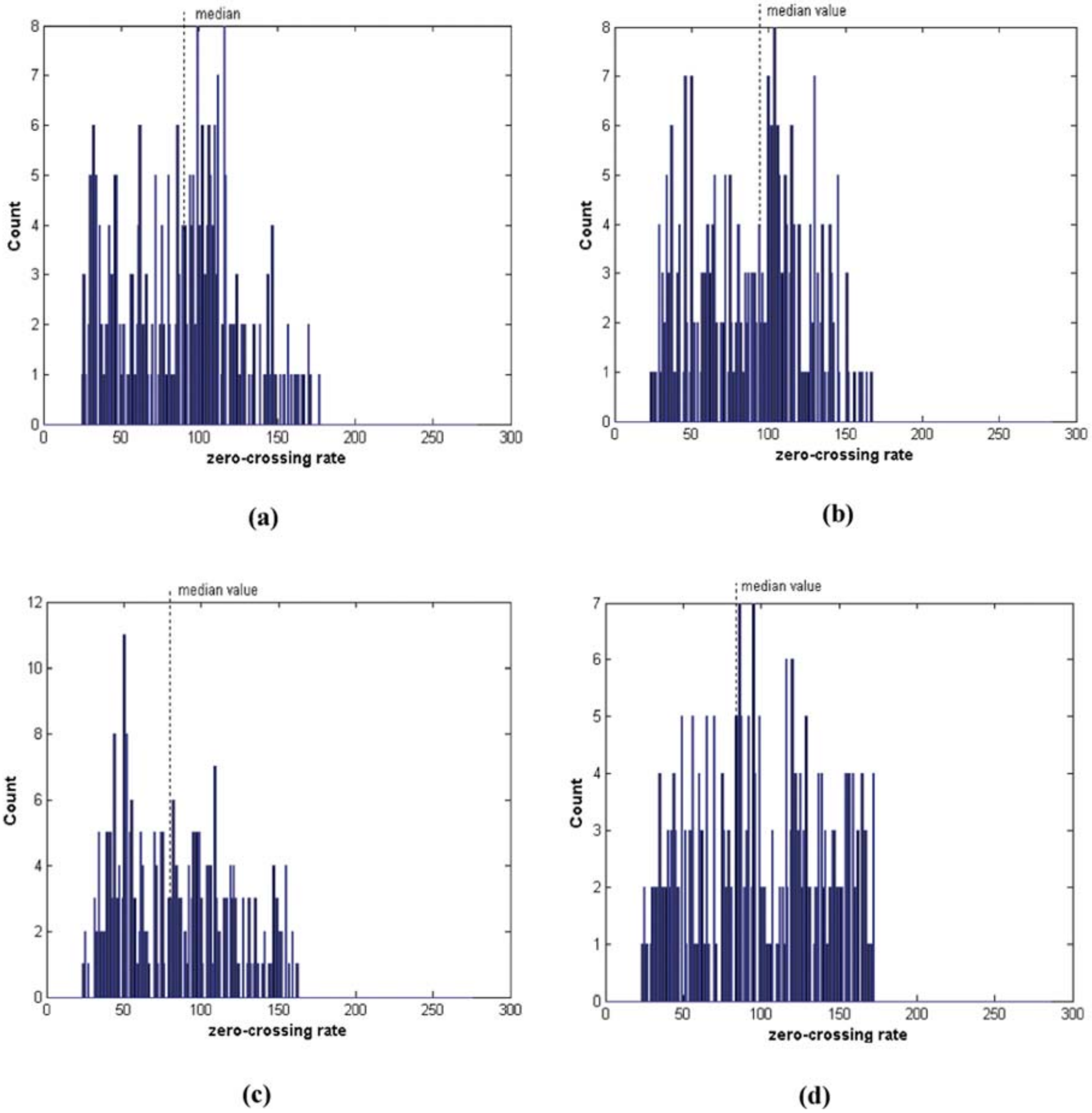
where *N* denotes the length, in samples, of the speech segment, $x_i(n)$. Our investigation has shown that, when the zero-crossing rate for a given speech segment exceeds a given threshold; the segment is likely to be unvoiced. However, our investigation has also shown that the distributions of the voiced and unvoiced segments in the speech inevitably overlap. This is demonstrated in Fig. (**3**), where histograms for the distribution of the zero-crossing rates for two different male speakers (a & b) and two different female speakers (c & d) are shown. In all cases, the length of speech signal used is 10 s. Hence, a threshold for discrimination between voiced and unvoiced speech using the zero-crossing rate needs to be carefully determined. Depending on the above-mentioned property and on an approach reported in [2], it has been decided to use a continually up-dated value of the median of the zero-crossing rates in this algorithm as a second threshold, *TH*2:

$$TH2 = median\{ZCR_m\}_{m=1}^{M} \tag{11}$$

where *M* is the index of current speech segment. Therefore, if the measured zero-crossing rate of the *i*th segment, ZCR_i, is equal or higher than *TH*2, the segment is most likely to be an unvoiced segment; otherwise, it is likely to be a voiced segment.

V/UV Classification Based on the Use of Two Features

Although each of the extracted features was found to provide sufficiently accurate indication for classification of V/UV segments, further investigation showed that higher classification accuracy is achieved by using a decision based on the two features together. Fig. (**4**) shows the wavelet-based per subband energy distribution and the corresponding zero-crossing rates for each segment of the same speech signal whose waveform is shown in Fig. (**2**), when (a) the signal is clean, and (b) the signal is degraded by additive white noise at SNR =10 dB. In both cases, values of each feature show total agreement with the criteria discussed in Sections 3.1 and 3.2 above. In addition, there is a clear correlation between the behaviour of the two features, providing strong basis for marking voiced and unvoiced regions within the signal. Based on the above, our algorithm uses a criterion involving a joint examination of the two extracted features, such that a segment indexed by *i* is declared unvoiced if, and only if, the following logical expression is satisfied



(a)

(b)

(c)

(d)

Fig. (3). Histograms of zero-crossing rates for utterances taken from two different female speakers (**a** & **b**), and two different female speakers (**c** & **d**).

$$\left[(R_i < TH1) \wedge (ZCR_i \geq TH2)\right] \Rightarrow (i \in UV) \qquad (12)$$

where “$\wedge$” denotes the logical *AND* operation and *UV* is the set of unvoiced indexes. If one or both features are not satisfied, the segment is declared voiced speech.

Logic Smoothing of Classification Data

Due to overlapping between voiced and unvoiced regions in natural speech, the above two-feature V/UV classification may, in some cases, yields incorrect results. This would usually occur when examining a segment which falls at the end of a voiced interval. Erroneous decisions may also be obtained for low-level voiced segments, voiced segments with rapid amplitude fluctuations, and segments contaminated with additive noise. In an automatic classification mechanism, the above problems may not be detected. Hence, a smoothing process has been added to the algorithm, as illustrated in Fig. (**1**), to correct logically obvious errors in the classified data. The process is based on the assumption that for short portions of speech, such as the speech segments used in this algorithm, it would be illogical to detect a single voice segment within a number of successive unvoiced segments. For example if the output of the classification decision is: …*UUUVUU*…, then smoothing is applied to correct this string to …*UUUUUU*, and so on. The same smoothing/correction is applied when a single voiced segment is detected within a number of successive unvoiced segments.

IMPLEMENTATION& PERFORMANCE EVALUATION OF PROPOSED ALGORITHM

The proposed speech classification algorithm has been implemented and experimentally tested used a large database comprising a wide variety of speech records taken from the TIMIT database. The speech signals, which cover different speakers and utterances, were ranging from 2 s to 10 s in length, and organized into 100 speech files corresponding to a total of 30000 speech segments with each, having duration of 25 ms. A 4-level dyadic DWT is used to decompose the input speech signal into 5 frequency subbands. Different wavelet filters were considered, however the algorithm was found to offer the best performance when a Haar filter is used. The signals were accurately labelled and their V/UV regions marked in association with the chosen segment length using a combination of visual inspection, listening and a dynamic pitch tracking algorithm [10]. This generated a labelled reference database for verification of the results obtained by the application of proposed algorithm.

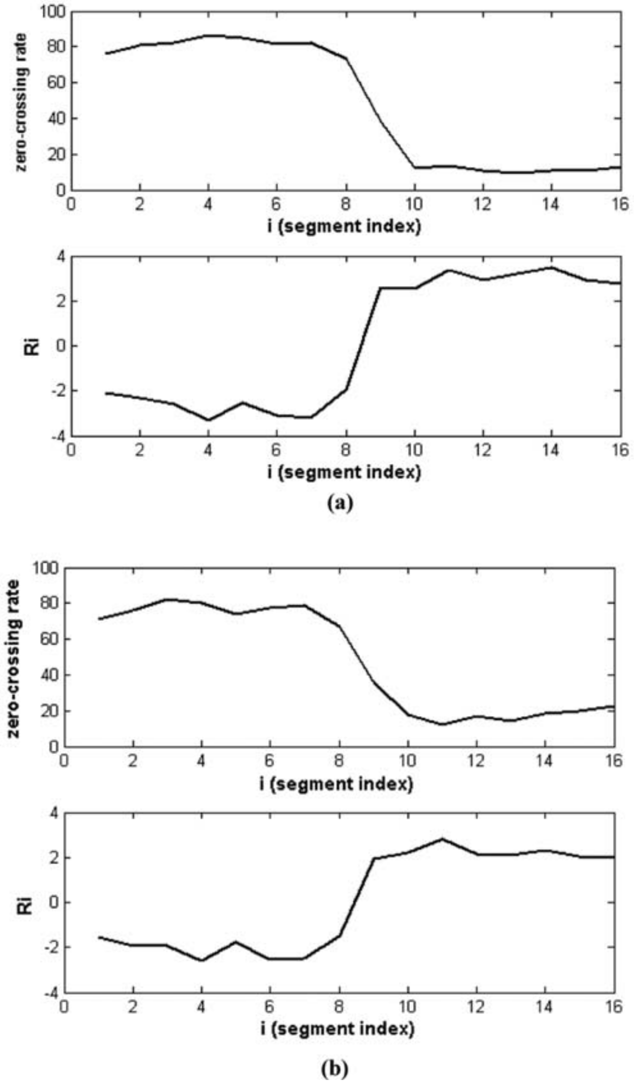


Fig. (4). The per band energy distribution and the zero-crossing rates feature for each segment of the speech signal used in Fig. **2**, when (**a**) the signal is clean, and (**b**) the signal is noise-degraded.

Using the algorithm, Fig. (**5**) shows the per segment distribution of the two extracted parameters for the case of a speech signal taken from a female speaker uttering the sentence: *“She had your dark suit in greasy-wash water all year”*, when the signal is (a) contaminated with noise at 15dB SNR, and (b) contaminated with noise at 5dB. The accuracy of the V/UV classification obtained by the algorithm was evaluated using an objective error measure, VuVER%, which represents the percentage of the total number of segments that have been erroneously classified. The measure covers both voiced-to-unvoiced and unvoiced-to-voiced error rates. The performance of the proposed algorithm has been evaluated under various noisy conditions. Tables **1** and **2** show the results of the analysis for male and

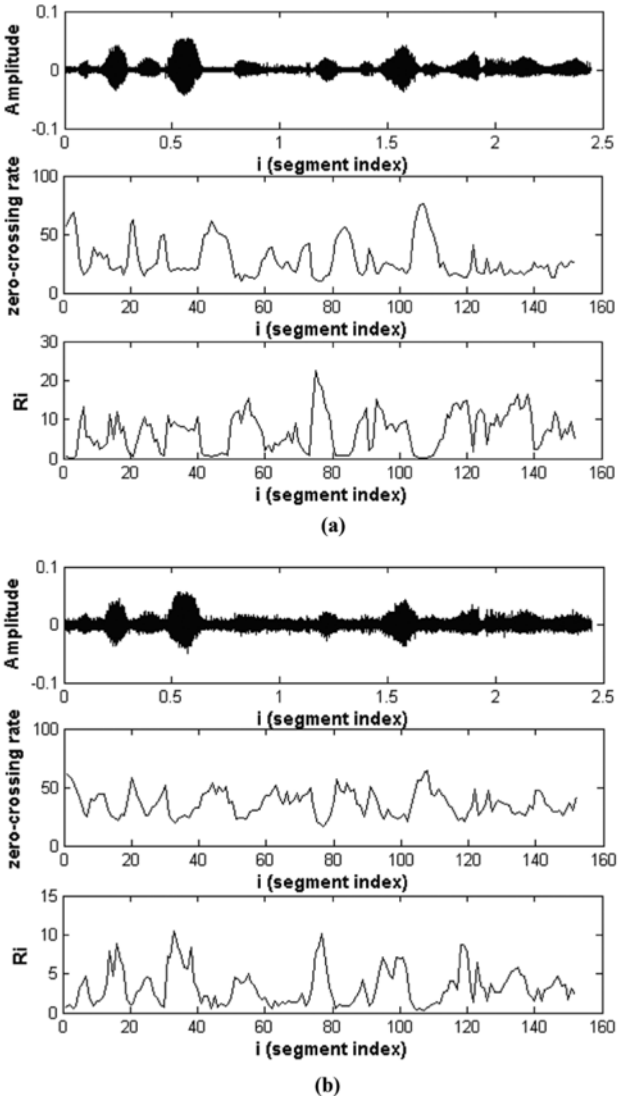


Fig. (5). The distributions of zero-crossing rates and per band average energy for a speech signal contaminated with noise at (**a**) SNR =15dB, and (**b**) SNR = 5dB.

Table 1. Performance of the Proposed Algorithm at Different SNRs for Male Speakers

	Vuver %		
SNR	**M1**	**M2**	**M3**
Clean	0.56	0.83	0.0
20dB	0.88	1.75	0.0
10dB	2.65	3.5	1.22
5dB	3.35	4.26	2.89
0dB	4.42	5.50	3.22

female speakers at different SNRs, generated by adding white Gaussian to the clean speech signals. It is clear that our algorithm performs well with both clean and noise-degraded speech. The results also show that for a highly noise-contaminated speech, better classification is achieved when dealing with male speakers. We believe that this phenomenon is due to the fact that female utterances usually contain shorter transient voiced segments as compared to male utterances. When contaminated with noise, the characteristics of such transient segments can be easily confused with those of unvoiced speech.

Table 2. Performance of the proposed algorithm at different SNRs for female speakers

	Vuver %		
SNR	**F1**	**F2**	**F3**
Clean	1.15	2.17	1.89
20dB	3.20	3.40	3.54
10dB	6.57	4.64	6.89
5dB	8.86	7.96	8.94
0dB	10.15	10.28	10.59

CONCLUSIONS

A new and novel speech classification algorithm was presented and its performance evaluated. The algorithm is based on extracting two features of the input speech: the ratio of the average energy in the wavelet low-subbands to that in the wavelet highest-subband, and the zero-crossing rate of each segment. By comparing the values of the extracted features to appropriately determined thresholds, the algorithm classifies the in the input speech into voiced and unvoiced regions. The performance of the algorithm was evaluated under different noisy conditions using a large database comprising a variety of speech signals from male and female speakers. Reported results have also shown that the proposed algorithm is robust to additive noise.

REFERENCES

[1] Qi. Y, Hunt BR. “Voiced–unvoiced–silence classification of speech using hybrid features and a network classifier”. *IEEE J. Trans.* Speech and Audio Processing; vol. 1, pp. 250-255, 1993

[2] Ahmadi S, Spanias AS. “Cepctrum-based pitch detection using a new statistical V/UV classification algorithm”. *IEEE J. Trans. Acoust., Speech and Audio Processing*; vol. 7; pp.333-338, 1999

[3] Rouat J, Liu YC, Morissette D. “A pitch determination and voiced / unvoiced decision algorithm for noisy speech”. *Speech Comm. J.* vol. 3, pp. 191-207, 1997

[4] Veprek P, Scordilis MS. “Analysis, enhancement and evaluation of five pitch determination technique”. *Speech Comm. J.* vol. 37 pp 249-270, 2002

[5] Rioul O, Vetterli M. “Wavelets signal processing”. *IEEE Signal Processing Magazin*; vol. 8, pp 14-38, 1991

[6] Donoho DL, Johnstone I M. “Ideal spatial adaptation by wavelet shrinkage”. *Biometrika*; vol. 81: pp 425-455, 1994

[7] Lu CT, Wang HC. “Speech enhancement using hybrid gain factor in critical-band-wavelet-packet transform”. *Digital Signal Processing J.* 2007; vol.17, pp 172-188, 2007

[8] Wang N, Zheng D-Z. “New algorithm for speech enhancement using node threshold wavelet packet transform”. *Chinese J. of Scient. Instr.*; vol 5, pp 952-955, 2007

[9] Wang ZL, Yang J, Zhang XW. “Combined discrete wavelet transform and wavelet packet decomposition for speech enhancement”. Proc. Intl. Conf. Sig. Process ICSP 2006; Vol 2.

[10] Wendt C, Petropulu AP. “Pitch determination and speech segmentation using the discrete wavelet transform”. Proc. IEEE Symp Circuits Systems ISCAS; 1996.vol. 2, pp. 45-48.

[11] Yumin Z, Yi Z, Ping L. “Improvement of pitch detection in noisy speech based on wavelet transform”. Proc. IET Intl. Conf .Wireless Mobile and Multimedia Networks ICWMMN, 2006, pp. 274-276.

[12] Janer L, Bonet JJ, Solano EL. “Pitch detection voiced/unvoiced decision algorithm based on wavelet transforms”. Proc. Intl. Conf .Spoken Language Process (ICSLP) 1996, 2: pp. 1209-1212.

[13] Kim JO, Hwang DJ, Paek HW, Chung CH. ‘An application of the merging algorithm with the discrete wavelet transform to extract valid speech–sound”. Proc. IEEE Workshop on Virtual and Intelligent Measurement Systems (VIMS) 2001; 67-70.

[14] Sheikhzadeh H, Abutalebi HR. “An Improved wavelet-based speech enhancement system”. Proc. Eurospeech 2001; pp. 120-124.

[15] Seok JW, Bae KS. “Speech enhancement with reduction of noise components in the wavelet domain”. Proc. IEEE Intl. Conf.Acoustic, Speech and Sig Process (ICASSP); vol. 2, pp. 1323-1326, 1997.

[16] Akansu AN, Haddad RA. “Multiresolution Signal Decomposition: Transforms, Subbands and Wavelets” 2^{nd} Ed., San Diego, USA. Academic Press, 2001.

Received: December 07, 2007 Revised: February 04, 2008 Accepted: March 03, 2008