Comparative Performance Study of Several Features for Voiced/ Non-Voiced Classification

Ykhlef Faycal[1] and Messaoud Bensebti[2]
[1]Multimedia Laboratory, Centre de Développement des Technologies Avancées, Algeria
[2]Electronics Department, University of Saad Dahlab, Algeria

Abstract: *This paper presents a comparative performance study of several time domain features for voiced/non-voiced classification of speech. Five classification schemes have been developed by combining one or two features amongst: Energy (E), Zeros Crossing Rate (ZCR), Autocorrelation Function (ACF), Average Magnitude Difference Function (AMDF), Weighted ACF (WACF), and the Discrete Wavelet Transform (DWT). The development of these classifiers was based on the selection of the lowest number of time domain features which allow voicing decision without the need of any frequency transformation or pre processing approaches. The performance of the classifiers has been evaluated on speech data extracted from the TIMIT database. Two different noise types: White and babble, taken from the NOISEX92 database have been incorporated to validate the developed classification schemes in noisy environments. An overall ranking of these classifiers for high and low Signal to Noise Ratios (SNRs) have been established based on the average value of the Percentage of classification accuracy (Pc).*

Keywords: *ACF, AMDF, DWT, E, non-voiced, voiced, ZCR, WACF.*

Received November 16, 2011; accepted January 28, 2013; published online April 4, 2013

1. Introduction

Accurate and reliable voiced/non-voiced classification of speech is a crucial pre-processing step in many speech processing applications and it is essential in most analysis and synthesis systems. The essence of classification is to determine whether the speech production system involves vibration of the vocal folds or not. For example, voicing determination is a crucial step in pitch detection problem. The accuracy of detection can significantly improve the performance of a pitch detector [1, 2].

Voiced speech consists of periodic or quasi periodic sounds made when there is a significant glottal activity (vibration of the vocal folds). Unvoiced speech is non periodic, random excitation sounds caused by air passing through a narrow constriction of the vocal track. Unvoiced sounds include the main classes of consonants which are voiceless fricatives, occlusives and stops.

When both quasi-periodic and random excitations are present simultaneously (mixed excitation, such as voiced fricatives), the speech is classified voiced because the vibration of vocal folds is part of the speech act. In other contexts, the mixed excitation could be treated by itself as a different class [7]. The non-voiced region includes silence and unvoiced speech [4].

A variety of techniques for robust voiced/ non-voiced classification have been reported in literature [3, 4, 7, 11]. The majority of them use hybrid approaches for voicing decision which include time or frequency domain features.

Several features have been used in the literature. We can mention, the Energy (E) of the signal, Zeros Crossing Rate (ZCR), Autocorrelation Function (ACF), Average Magnitude Difference Function (AMDF), Weighted ACF (WACF), Cepstral Function (CEP), Discrete Wavelet Transform (DWT) coefficients, first coefficient of a p^{th} order linear prediction analysis and harmonic measure [3, 4, 10]. The combination of evidence from multiple features can be done by using statistical models such as neural network, Gaussian Mixture Model (GMM) or hidden Markov model [3]. The combination of features can significantly offer an accurate classification which basically depends on the number of features incorporated in the model. On the other hand, the need for hardware implementation and real time applications requests the reduction of the features number which aims at decreasing the computational complexity.

The performance of these classifiers in terms of percentage of classification accuracy in noisy environments depends on the choice of the suitable feature and on some pre and post processing approaches (such as clipping, resampling and smoothing). The time domain acoustical features are widely used in practical implementations. They are generally preferred due to their low computation and precise estimation.

The main purpose of this paper is to study separately the performance of several time domain features for voiced/non-voiced classification of speech in clean and noisy environment and to establish an overall ranking of these features.

To achieve our purpose, five classification schemes that use only one or two features have been developed without the need of pre- or post processing stages. The classifiers are given as follows:

1. ACF.
2. AMDF.
3. WACF.
4. ZCR. E.
5. DWT. E.

Manually segmented speech signals from TIMIT data base [5] are used to measure the success of the classification into voiced and non-voiced frames.

The performance of the developed classifiers is evaluated by using an additive white and babble noises, extracted from the NOISEX 92 data base [13]. Different Signal to Noise Ratios (SNRs) of the input signal have been used. They are ranged from 30 to -5dB.

Our paper is organized as follows: In section 2 the detailed implementations of the five voiced/non-voiced classifiers are reviewed. In section 3, the performance evaluation of the classifiers is given. Section 4 gives a conclusion and future perspectives.

2. Voiced/ Non-Voiced Classifiers

Five distinct classification schemes for voiced/ non-voiced decision were investigated. The development of these classifiers was based on the selection of the lowest number of time domain features without the need of any frequency transformation. In the following, a detailed explanation of each classifier is provided.

2.1. ACF

The fist classifier is based on one acoustical feature which is the ACF. Figure 1 ($i = 1$) shows a block diagram of the classification scheme. The approach is based on frame by frame processing of the speech signal using a stationary rectangular window of 22.5ms duration. The used speech signal has a sampling frequency (F_s) of 16kHz. The frame $x(n)$ is low pass filtered to 700Hz (A 20-point linear phase, Finite Impulse Response (FIR) digital filter) in order to eliminate the formant structure of the speech signal. The first stage of the processing is the computation of the ACF using equation 1:

$$\varphi_1(\tau) = \frac{1}{N}\sum_{n=0}^{N-\tau-1} x(n)\,x(n+\tau) \qquad (1)$$

$$\tau = n_1 : n_2$$

Where N, is the length of the frame which is equal to 360. τ is a lag number. n_1 and n_2 represent the rang of ACF computation which corresponds to the frequency band of pitch [70 to 600Hz], $n_1 = 26$, $n_2 = 200$.

The $\phi_1(\tau)$ is characterized by a large peak for voiced frames which decreases for non-voiced frames.

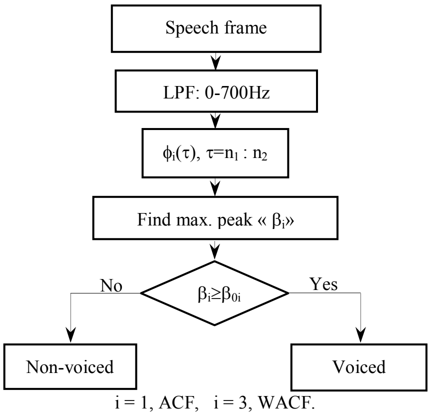


Figure 1. Block diagram of the correlation classifiers.

Theoretically, the ACF of unvoiced frames are not characterized by apparent peaks. However, the ACF of silence may include some peaks due to the spectral shape of silence regions which are comparable to voiced ones. Basically, they are distinguished by the peaks of their ACF (The largest peak of the ACF is very small for silence compared to voiced frames). The second stage of the processing is the extraction of the largest peak in the ACF which is called β_1. Then, in the third stage, this peak is compared to a constant threshold β_{01}. If β_1 is greater than β_{01}, the frame is classified voiced; otherwise, it is classified non-voiced. It should be noticed that it is possible to use a normalized version of the autocorrelation function for voicing decision. In this case, we would be obliged to add a silence detector in order to eliminate the wrong decisions (silence frames classified as voiced ones) as presented in [8]. A non-normalized version of the ACF (with constant threshold) has been used in this study in order to reduce the number of features in the classification (no need for a silence detector for voiced/non-voiced classification).

2.2. AMDF

Figure 2 shows a block diagram of the classification based on the AMDF feature. It is clear that the classifier follows the same steps as the previous one, except for the function which is the AMDF given by equation 2:

$$\varphi_2(\tau) = \frac{1}{N}\sum_{n=0}^{N-\tau-1} \left| x(n) - x(n+\tau) \right| \qquad (2)$$

$$\tau = n_1 : n_2$$

Where N, is the length of the frame which is equal to 360 and τ is a lag number.

n_1 and n_2 represent the range of AMDF computation which corresponds to the frequency band of pitch [70 to 600Hz], $n_1 = 26$, $n_2 = 200$.

The AMDF is a variation of the ACF where instead of correlating the input speech at various delays where multiplications and summations are formed at each value, a difference signal is formed between the delayed speech and the original, and at each delay value, the absolute magnitude is taken. Unlike, ACF, however, the AMDF calculations require no multiplications, a desirable propriety of hardware and real time applications [14].

The $\phi_2\left(\tau\right)$ is characterized by several valleys which appear periodically for voiced speech. The global minimum valley point β_2 is used in the decision. If it is lower or equal than a constant threshold β_{02}, the speech is classified voiced, otherwise, it is classified non-voiced. The same as ACF classifier, only AMDF function is used in the classification process, and there is no need to detect silence regions.

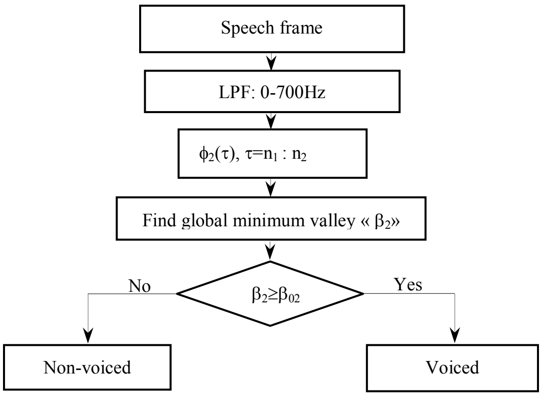


Figure 2. Block diagram of the AMDF classifier.

2.3. WACF

The WACF has been firstly proposed by Shimamura and Kobayashi [12] for pitch detection. Utilizing that the AMDF has similar characteristics to the ACF, the ACF is weighted by the reciprocal of the AMDF to form what is called WACF. It has been shown in [12] that this function has good performance for pitch detection in white noise environment. The proposed approach for voicing decision is shown in the diagram block of Figure 1 ($i = 3$), where $\phi_3\left(\tau\right)$ denotes the weighted ACF given by the following equation:

$$\varphi_3\left(\tau\right) = \frac{\varphi_1\left(\tau\right)}{\left(\varphi_2\left(\tau\right)+\kappa\right)} \qquad (3)$$

$$\tau = n_1 : n_2$$

Where τ is a lag number,

$\phi_1\left(\tau\right)$ is the ACF,

$\phi_2\left(\tau\right)$ is the AMDF,

n_1 and n_2 represent the rang of WACF computation which correspond to the frequency band of pitch [70 to 600Hz], $n_1 = 26$, $n_2 = 200$.

κ is a constant number used to avoid divergence of the directly inversed $\phi_2\left(\tau\right)$ at lag zero (because $\phi_2\left(0\right) = 0$). In this study, this number is not significant since $\phi_2\left(\tau\right)$ is computed between n_1 and n_2. It is set to 0.1.

The voicing decision follows the same steps as in the ACF classifier. The largest peak β_3 is compared this time to the β_{03} threshold.

2.4. ZCR. E

One of the simplest ways to perform voiced/non-voiced classification is based on the use of the ZCR and E of speech signal [9]. The ZCR is a measure of number of times in a given time interval in which the amplitude of speech signals passes through a value of zero.

The F_s determines the time resolution of zeros crossing rates. In our study, F_s is set to 16kHz. The *ZCR* can be computed by using the following equation:

$$ZCR = \sum_{n=1}^{N}\frac{1}{2}\left|sgn\left(x\left(n\right)\right)-sgn\left(x\left(n-1\right)\right)\right| \qquad (4)$$

Where $x\ (n)$ is the speech frame, N is the frame length, sgn is a function which gives the sign of the sample. It takes 0 for null values, -1 for negative ones and 1 for positive values.

Basically, unvoised speech exhibits a higher ZCR than voiced speech or silence. Therefore, the voiced/silence speech distinction needs another feature to accomplish a correct voiced/non-voided decision. Generally, the *E* of the speech is used as second feature. It provides a convenient representation that reflects the variation of the amplitude of a speech frame. The average energy of a given silence frame is much lower than that of voiced and unvoiced ones. Furthermore, it can be observed that the *E* of voiced frames can be higher than unvoiced ones. The average energy is computed as follows:

$$E = \frac{1}{N}\sum_{n=0}^{N-1} x\left(n\right)^2 \qquad (5)$$

Figure 3 shows the block diagram of the developed classifier. After frame acquisition using a stationary rectangular window of 22.5 ms duration (N = 360), the average energy of the frame is computed. Then, this latter is compared to a constant threshold E_0. If the computed energy E is lower or equal to E_0, then the frame is classified non-voiced. Otherwise, the *ZCR* is computed and compared to a constant threshold ZCR_0. If *ZCR* is lower or equal to ZCR_0, the speech frame is classified voiced; if not, it is classified non-voiced.

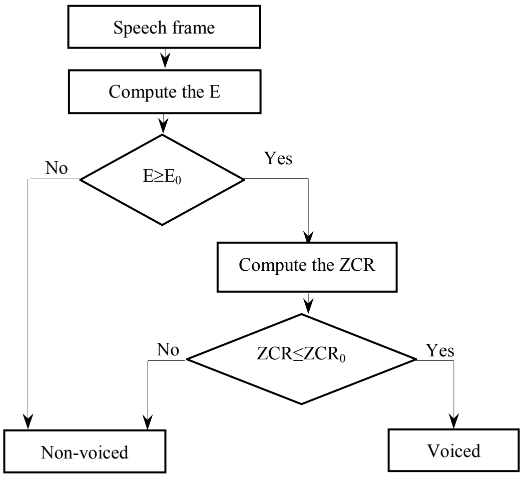


Figure 3. Block diagram of ZCR.E classifier.

2.5. DWT. E

Several approaches for voicing decision that use the DWT are reported in the literature. For example, we have the Johnson algorithm [6], which is able to classify the speech signal into three main categories: voiced, unvoiced and mixed speech. Basically, the Johnson algorithm, does not detect silence regions. Thus, in this study, a modified version of the previous algorithm is proposed to perform voiced/non-voiced classification. The modification is achieved by using the average energy as second feature.

In the classification process as shown in Figure 4, the signal is fragmented into frames of 22.5ms duration by using a rectangular window. The algorithm starts by computing the average E of the frame using equation 5.

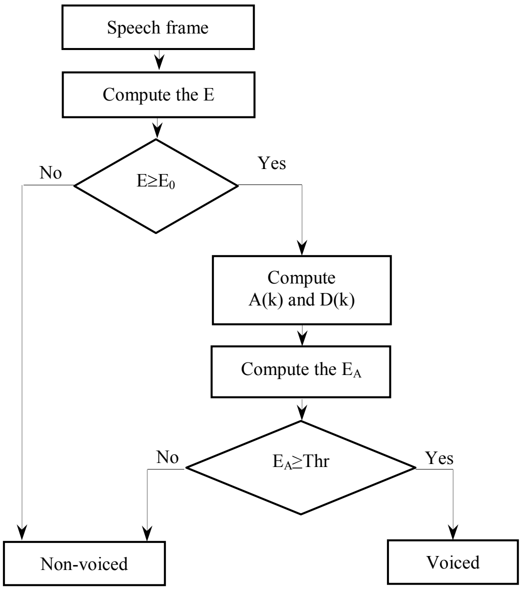


Figure 4. Block diagram of DWT. E classifier.

Assuming that, the discrete samples of the signal are normalized between -1 and 1. The frame is classified silence if its corresponding E does not exceed a constant threshold E_0. If it is not, the approximation (A (k)) and detail (D (k)) coefficients are computed using Haar Wavelet as follows:

$$A(k) = \sum_{\ell=-\infty}^{\infty} x(\ell)\, h_0(2k-\ell) \tag{6}$$

$$D(k) = \sum_{\ell=-\infty}^{\infty} x(\ell)\, h_1(2k-\ell) \tag{7}$$

$$k = 0 : S-1$$

Where x (n) is the speech frame with length N. h_0 (n) is the low pass filter associated to the Haar Wavelet. h_1 (n) is the associated high pass filter and M is its length. $S=\frac{N+M-1}{2}$ is the length of the Wavelet coefficients:

$$h_0(n) = \begin{cases} \frac{1}{\sqrt{2}} & if\ n \in \{0,1\} \\ 0 & else \end{cases} \tag{8}$$

$$h_1(n) = (-1)^n\, h_0(n) \tag{9}$$

In the second step, the energies of level one approximation (E_{app}) and detail (E_{det}) coefficients are computed and normalized by the sum of both energies.

Considering that the global energy represents 100%. If the percentage of the normalized energy concentrated in level one approximation coefficients (E_A) is less than a constant threshold Thr, then, the segment is classified non-voiced, otherwise, it is classified voiced.

Several solutions have been reported in the literature to find the optimal thresholds used in the proposed classifiers. For instance, we have: the average value, the median value or a constant percentage of a set of values which correspond to each feature in the utterance [1].

The optimal thresholds selection used in this study will be discussed in the performance evaluation section.

3. Performance Evaluation

3.1. Criteria of the Test

The performance of the five classifiers were tested on speech database which was hand labeled into voiced/non-voiced regions. Three measures of performance were used [1, 4]:

1. Voiced speech classified as non-voiced (VNV error).
2. Non-voiced speech classified as voiced (NVV error).
3. Percentage of classification accuracy (Pc):

$$P_c = 1-(A \times VNV + B \times NVV) \tag{10}$$

Where A and B represent the percentage of voiced and non-voiced frames in the speech utterance, respectively.

The two types of error listed above occurred during the initial speech classification into voiced and non-voiced regions, where non-voiced speech was considered as voiced or voiced speech misclassified as non-voiced.

3.2. Speech Data

The developed classifiers are evaluated on the TIMIT database [5]. Speech underwent extensive manual labeling before it could be used.

The spoken material consists of a set of 26 rich English sentences from TIMIT database (13 female and 13 male) that contain several dialects and acoustic forms (weak voiced speech, rapid voiced non-voiced transition).

The percentage of voiced speech samples in each of the utterance is maintained at 50% (A and B equal to 0.5 in equation 10) by appending required duration of silence. Two experienced persons performed the manual classification of the spoken material. The original time waveform was used as the primary tool, with spectrograph utilized only in few cases where the waveform was insufficient to make the decision.

3.3. Performance of the Classifiers

The classification schemes developed in this paper are defined as threshold-based classifiers. The classification performance is directly related to the optimal choice of the thresholds given as follows:

- β_{01} for ACF.
- β_{02} for AMDF.
- β_{03} for WACF.
- E_0 and ZCR_0 for ZCR. E.
- E_0 and *Thr* for DWT. E.

The main purpose of this study is to assess each time domain feature by using an optimal decision level for voicing classification of English. Consequently, an optimal threshold for each feature is needed to evaluate the global performance of the classifiers.

A practical approach is to seek a value that gives the optimal classification for each utterance in clean environment and then, to use it in noisy environment. The optimal thresholds are obtained after the computation of the requested feature for each frame. The median value is taken as threshold. This procedure is used to extract: β_{01}, β_{02}, β_{03} and ZCR_0 which are respectively the median values of β_1, β_2, β_3 and ZCR. This computation is performed for the entire frames of each utterance. For instance, β_{01} is obtained by calculating the median value of a set of β_1 values calculated for each frame in an utterance. It is updated for every utterance in the database. The same procedure is performed for the remained features.

The silence threshold (E_0) used in ZCR. E and DWT. E classifiers, is empirically set to 0.05 which is considered as silence decision level for a frame length of 22.5ms. The *Thr* threshold used in the DWT. E classifier is set to 77%.

The performance of the classifiers is reported in Table 1 for clean and noisy speech obtained at different SNRs of the input signal. The White and babble noises extracted from the NOISEX92 database [13] have been incorporated in the experiment.

The entire classifiers have good performances in clean environment which are degraded according to the type and SNRs of the added noises. The WACF has the best performance in white noise environment. The noticed degradation is essentially related to the NVV errors (in low SNR levels) which increase by the diminution of the SNRs. The ACF has comparable performance to the previous classifier especially for SNRs higher or equal to 5dB. A serious degradation is noticed at lower SNRs.

The ZCR. E and AMDF classifiers are seriously influenced by the addition of white noise. The Pc of the AMDF is reduced due to the detected high NVV errors. However, for the ZCR. E classifier, the degradation in the accuracy is related to the VNV errors.

The ZCR. E performs better than the AMDF at high and low SNRs (30dB, 0 and -5dB). The performances are reversed for medium SNRs (20, 10 and 5dBs). The DWT. E classifier has a uniform performance evaluation, and assures better accuracy especially at low SNRs.

The same as ZCR. E classifier, the Pc of the DWT. E is directly related to VNV errors especially at low SNRs.

Table 1. Performance of voiced/non-voiced classification.

	SNR		ACF	AMDF	WACF	ZCR.E	DWT.E
White noise	Clean	Pc	97.58	95.92	97.66	97.12	95.90
		VNV	2.20	2.93	2.37	3.23	0.99
		NVV	2.62	5.23	2.31	2.52	7.20
	30dB	Pc	97.11	95.85	97.62	96.49	95.87
		VNV	3.28	2.89	2.74	5.75	0.94
		NVV	2.49	5.39	2.01	1.26	7.30
	20dB	Pc	97.05	95.63	97.60	91.51	95.53
		VNV	3.43	2.44	2.86	16.74	1.31
		NVV	2.47	6.30	1.93	0.24	7.62
	10dB	Pc	96.52	83.90	96.57	73.55	93.04
		VNV	3.60	0.28	5.25	51.90	12.92
		NVV	3.34	31.91	1.60	1.00	0.99
	5dB	Pc	93.97	62.84	95.28	62.55	84.75
		VNV	3.24	0.04	6.21	73.91	30.22
		NVV	8.81	74.28	3.21	0.99	0.27
	0dB	Pc	63.74	50.00	87.29	54.22	68.92
		VNV	0.21	00.00	7.71	90.57	62.15
		NVV	72.31	100.0	17.69	0.97	00.00
	-5dB	Pc	50.75	50.00	67.08	51.18	56.19
		VNV	0.04	00.00	4.55	96.66	87.61
		NVV	98.45	100.0	61.28	0.97	00.00
Babble	30dB	Pc	96.92	95.574	97.46	96.02	95.56
		VNV	3.28	2.74	3.23	3.72	1.02
		NVV	2.87	6.11	1.84	4.23	7.84
	20dB	Pc	95.41	80.92	94.34	85.97	84.90
		VNV	2.92	1.08	5.55	2.65	0.58
		NVV	6.25	37.07	5.75	25.40	29.60
	10dB	Pc	54.92	50.28	68.95	57.25	55.46
		VNV	0.25	00.00	4.95	1.11	0.20
		NVV	89.89	99.42	57.14	84.38	88.86
	5dB	Pc	50.41	50.00	55.64	56.10	53.34
		VNV	00.00	00.00	2.80	0.77	0.07
		NVV	99.18	100.0	85.91	87.01	93.23
	0dB	Pc	50.00	50.00	50.75	54.17	51.60
		VNV	00.00	00.00	0.81	0.67	0.04
		NVV	100.0	100.0	97.70	90.98	96.76
	-5dB	Pc	50.00	50.00	50.74	52.18	50.56
		VNV	00.00	00.00	0.81	0.55	00.00
		NVV	100.0	100.0	97.69	95.07	98.87

In babble noise environment, the WACF classifier has the best Pc at high SNRs (30, 20, and 10dBs). However, the ZCR. E has better performance at low SNRs (5, 0 and -5dB) among the studied classification schemes. The other classifiers have variable performances which depend on the used SNRs. It can be noticed that the main source of errors in babble noise environment is related to NVV errors. This remark is valid for the five classifiers.

The overall ranking of the classification schemes is not a simple operation. The performance can show a discrepancy for each noise type and SNR level.

In this study, the overall ranking is established for high and low SNRs based on the average value of Pcs respectively at (30, 20, 10 dB) and (5, 0, -5 dB). The results are reported in Table 2.

In white noise environment, the WACF classifier ranks first with overall Pcs of 97.26 and 83.22% for high and low SNRs, respectively. The ACF ranks second at high SNRs followed, respectively by the DWT. E and AMDF. The ZCR. E remains in the last position. For low SNRs, the DWT. E classifier ranks second followed respectively by the ACF and ZCR. E. The AMDF classifier ranks last.

In babble environment, the WAFC is constantly in the first rank (Pc = 86.92%) for high SNRs followed by the ACF classifier. The ZCR. E and DWT. E classifiers rank, respectively in the third and fourth position. The AMDF remains at the last rank. For low SNRs, the ZCR. E ranks first with overall Pcs of 54.15%. The WACF ranks second with a Pc lowered by 1.77%. The remained positions are for DWT. E, ACF and AMDF, respectively.

Table 2. Average values of the percentage of classification accuracy.

	SNR*	ACF	AMDF	WACF	ZCR. E	DWT. E
White noise	High	96.89	91.79	97.26	87.18	94.81
	Low	69.48	54.28	83.22	55.98	69.95
Babble	High	82.42	75.59	86.92	79.75	78.64
	Low	50.14	50.00	52.38	54.15	51.83

*High SNR: Average value of Pcs at (30, 20 and 10 dB)
Low SNR: Average value of Pcs at (5, 0 and -5 dB)

4. Conclusions

This paper reported the results of performance evaluation of five voiced/non-voiced classification schemes which use one or two features without any pre or post processing approaches.

Based on a variety of error measurements, the performances of the studied classifiers for different noise environments (white and babble noise at high and low SNRs) were highlighted. It has been noticed that the degradation of the percentage of classification accuracy of the five classification schemes is proportionally related to the SNR level. While the performance degradation in white noise environment of ZCR. E and DWT. E is related to VNV errors, the ACF, AMDF and WACF performance is directly related to NVV errors in the same environment. On the other hand, in babble noise environment, the performance degradation of the entire classifiers is directly related to the NVV errors. The overall ranking of the classification schemes is not a trivial problem; it depends on the application and the environment of the speech signal. In this study, the ranking of the classifiers was established based on the average values of the percentage of classification accuracy for each noise type. In white noise environment, the WACF ranks first for high and low SNRs. Conversely, in babble noise environment, the ZCR. E ranks first at low SNRs, and the WACF ranks first at high SNRs.

The combination of correlation features (ACF, AMDF and WACF) with ZCR or DWT approximation coefficients will significantly improve the accuracy and reduce the VNV and NVV errors. However, the computation complexity will increase.

In future works, the performance of the studied classification schemes will be evaluated in other noise types. Further, a two by two feature combination will be investigated for better accuracy.

Acknowledgements

Authors would like to thank Y. Boudjahfa and D. Belabdi post graduate English students, respectively at Batna University and ENS school of Algiers for their helpful suggestions and language support.

References

[1] Ahmadi S. and Spanias A., "Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333-338, 1999.

[2] Allam M. "Speech Segmentation in Synthesized Speech Morphing Using Pitch Shifting," *the International Arab Journal of Information Technology*, vol. 8, no. 2, pp. 221-226, 2011.

[3] Atal B. and Rabiner L., "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201-212, 1976.

[4] Dhananjaya N. and Yegnanarayana B., "Voiced/Nonvoiced Detection Based on Robustness of Voiced Epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273-276, 2010.

[5] Garofalo J., Lamel L., Fisher W., Fiscus J., Pallett D., Dahlgren N., and Zue V., "TIMIT Acoustic-Phonetic Continuous Speech Corpus Linguistic Data Consortium," *Linguistic Data Consortium*, Philadelphia, Trustees of the University of Pennsylvania, 1993.

[6] Johnson J., "Discrete Wavelet Transform Techniques in Speech Processing," *in Proceedings of IEEE TENCON, Digital Signal Processing Applications*, Australia, pp. 514-519, 1996.

[7] Qi Y. and Hunt B., "Voiced-Unvoiced-Silence Classification of Speech Using Hybrid Features and a Network Classifier," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 250-255, 1993.

[8] Rabiner L., Cheng M., Rosenberg A., and McGonegal C., "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Transactions on Acoustics, Speech, and*

Signal Processing, vol. 24, no. 5, pp. 399-418, 1976.

[9] Rabiner L. and Schafer R., *Digital Processing of Speech Signal*, Prentice-Hall, Englewood Cliffs, 1978.

[10] Routray A., Kabisatpathy P., and Mohanty M., "A Statistical Approach for Voiced Speech Detection," *Special Issue of International Journal of Computer and Communication Technology*, vol. 2, no. 2-4, pp. 52-55, 2010.

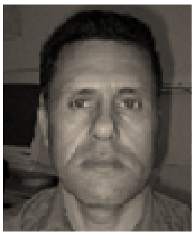
[11] Shahnaz C., Zhu W., and Omair M., "An Approach for Voiced/Unvoiced Decision of Colored Noise-Corrupted Speech," *in Proceedings of IEEE International Symposium on Circuits and System*s, USA, pp. 3944-3947, 2007.

[12] Shimamura T. and Kobayashi H., "Weighted Autocorrelation for Pitch Extraction of Noisy Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 9, no. 7, pp. 727-730, 2001.

[13] Varga A. and Steeneken H., "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and An Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Communication*, vol. 12, no. 3, pp. 247-251, 1993.

[14] Yu-Min Z., Zhen-Yang W., Hai-Bin L., and Lin Z., "Modified AMDF Pitch Detection Algorithm," *in Proceedings of International Conference on Machine Learning and Cybernetics*, Xian, China, vol. 1, pp. 470-473, 2003.

Ykhlef Fayçal graduated from the Saad Dahlab University of Blida, Algeria, in 2002 in electronic engineering. He received his Master of Science in speech and image processing from the same university in 2005. At present, he is a researcher at the Multimedia Laboratory of "Centre de Développement des Technologies Avancées", Algiers, Algeria. His research interests include medical applications, signal and speech processing.



Messaoud Bensebti graduated from the Ecole Nationale Polytechnique of Algiers, Algeria, in 1987 in electrical engineering. He received his PhD in electronics from Bristol University, UK in 1992. At present he is a professor in the Electronics Department of Saad Dahlab University of Blida, Algeria. His current research activities are concentrated in the areas of mobile radio broad band communication systems and signal processing.