A Review Paper on Voiced/Unvoiced Classification

Bhumika Nirmalkar[1] Dr. Sandeep Kumar[2]
[2]Assistant Professor
[1,2]Department of Electronics and Telecommunication
[1,2]Rungta College of Engineering and Technology Bhilai, India

Abstract— This paper presents voiced/unvoiced classification based on various methods like empirical mode decomposition, wavelet, cepstrum, zero crossing rate, short time energy etc and statistical model such as neural network, hidden Markov model (HMM) and Gaussian mixture model (GMM). In most of these techniques the voiced/unvoiced classification is usually perform by means of placing a threshold value with a few acoustic functions say, short time energy, zero crossing rate and so on. The primary trouble is the purpose of effective threshold which impact the type issues performance is likewise comes to a decision the choice of threshold. We have observed that a hybrid technique for the voiced/unvoiced classification can enhance the performance of the existing schemes.

Key words: Voiced/unvoiced classification, Cepstrum, EMD

I. INTRODUCTION

Relaible classification of short time speech signal into voiced and unvoiced in many speech processing applications like language identification[1],multi-rate speech coder[2,3], speech signal modeling[4] and there are some prominent speech signal application like pitch frequency estimation[5,6], identification of the glottal closure instants(GCIs)[7],which require knowledge of only the voiced regions of the speech signal. Speech is an acoustic signal produced from a speech production system. Producing speech sounds, the air flow from your lungs first passes the glottis and then your throat and mouth. Depending on speech sound the speech production can be broadly categorized into three activities:-Voiced speech, unvoiced speech, Silence region. If the voiced signal, input excitation is nearly periodic impulse sequence, then the corresponding speech looks visually nearly periodic. During the production of voiced speech; the air exhaling out of lungs through the trachea is interrupted periodically by the vibrating vocal folds. Due to this, the glottal wave is generated that excites the speech production system resulting in the voiced speech. . If the unvoiced signal, input excitation is random noise-like, then the resulting speech will also be random noise-like without any periodic nature. During the production of unvoiced speech, the air exhaling out of lungs through the trachea is not interrupted by the vibrating vocal folds. If the silence region, there is no excitation supplied to the vocal tract and hence no speech output. Many algorithms have been reported to solve voiced/nonvoiced (VnV) classification problems [8]. The discrimination is usually performed by setting a threshold value with some acoustic features say, short-term energy, zero-crossing rate, etc. The main problem of these methods is the determination of an effective threshold which affects the classification performance. Various method for classification of voiced/unvoiced like empirical mode decomposition, wavelet, and cepstrum based pitch detection, fuzzy logic etc.

The rest of the paper is organized as follows. In Section II, a description of the implementation of enhanced cepstrum based voiced/unvoiced speech classifier is given. In Section III, the results of the performance analysis are presented. Concluding remarks are given in Section IV

II. LITERATURE REVIEW

Various types of work have been done in the field of voiced/unvoiced classification and Various Method Were Adopted for Voiced/Unvoiced Classification. Molla et.al [9] proposed "voiced/unvoiced speech classification using adaptive thresholding with bivariate EMD". They proposed an effective method of voiced/unvoiced classification without any use of training data and prior knowledge, to achieve robust and data adaptive voiced/unvoiced classification technique which is suitable for real world speech processing application. They found that classification efficiency was better than that of the recently reported algorithms.

Abhay Upadhyay and Ram Bilas Pachori [10] proposed "instantaneous detection of voiced/non-voiced detection based on the method variation mode decomposition (VMD)". This method does not require prior information of the pitch. They found that it's provided better accuracy and performance of voiced and unvoiced classification.

Vinayak Abrol et.al [11] proposed "voiced/nonvoiced classification in compressively sensed speech signals". This method is based on compressive sensing (CS)/Sparse coding for detection of voiced/nonvoiced classification. They showed the sparse vector contains the source characteristics of the speech signals, if a suitable dictionary is chosen. Using an information theoretic based criterion the behaviour of sparse vector is quantified. An adaptive threshold selection scheme used for final voiced/nonvoiced classification. They found this method selected (voiced) region, which can be used for application of speaker verification.

Mohammaed Algebraic et.al [12] proposed "voice and unvoiced classification using fuzzy logic". This algorithm is based on features Zero crossing rate, Short time energy for classification of voice,unvoiced and silence. They showed this method successfully classified the speech signals.

Mollaet.al [13] proposed "Adaptive thresholding approach for robust voiced/unvoiced classification". They introduced a robust voiced/unvoiced classification method by using linear model of empirical mode decomposition (EMD) controlled by Hurst exponent. They showed This algorithm improves the classification performance.

Pooja jain and Ram Bilas Pachori [14] proposed " a pseudo Wigner-Ville distribution (PWED based method) for the voiced/unvoiced detection in noisy speech signals". She obtained marginal energy density with respect to time (MEDT) which is used as a feature to provide

voiced/unvoiced classification and allowed instantaneous detection of voiced regions. Also this method does not require knowledge of pitch frequency. They found that the performance of algorithm was improved for clean and noisy signals.

Poonam Sharma and Abha Kiran Rajpoot [15] proposed "identifying the voice,unvoiced and silence chunks in speech". This algorithm is based on Zero crossing rate, Short time energy, and fundamental frequency for identifying the speech signals. They found better accuracy and data collected four different speakers.

Senturk Zekeriya et.al [16] proposed "Voiced-unvoiced classification of speech using autocorrelation matrix". their method is based on, signal energy, the peak-to-peak difference of the autocorrelation function, number of zero crossings of the autocorrelation function and the unit delay autocorrelation coefficient all together. The accuracy of the proposed method found 100% for women and 98% for men.

Ykhlef Faycal and Messaoud Bensebti [17] proposed "a comparative performance study of several time domain features for voiced/unvoiced classification of speech". They have considered five classification schemes based on Energy (E), Zero crossing rate (ZCR), Autocorrelation Function (ACF), Average Magnitude Difference Function (AMDF), Weighted ACF (WACF) and Discrete Wavelat Function (DWT) for their study. They evaluated the performance of five voiced/unvoiced classification scheme one or two features without any pre or post processing approaches.

Mojtaba Radmard et.al [18] proposed "a new method of voiced/unvoiced classification based on clustering". Their algorithm is based on analysis of cepstral peak, zero crossing rate, and autocorrelation function (ACF) peak of short-time segments of the speech signal by using some clustering methods The advantage of this clustering based method is getting rid of determining a threshold. So it is highly speaker independent. They showed better satisfactory performance for identification of voiced and unvoiced segments of speech.

Dhananjaya, N., and B. Yegnanarayana [19] proposed "a new method for voiced/unvoiced classification based on epoch extraction". This method uses zero frequency filtered speech signal is used to extract the epochs. Features of this method are depends on excitation source information. They found that this method was better than the normalized cross correlation based voiced/unvoiced classification.

Ekaterina Verteletskaya et.al [20] proposed "pitch detection algorithms and voiced/unvoiced classification for noisy speech". Their algorithm is based on cepstral analysis; time auto co-relation, spectral temporal auto co-relation (STA) and average magnitude difference function. He found that all of the algorithms of voiced/unvoiced classification give good performance for clean speech.

Md. Khademul Islam Molla et.al [21] proposed "robust voiced/unvoiced speech classification using empirical mode decomposition and periodic correlation model". His method analyzes the signal by nonlinear and non stationary signal which is used as a filter for additive noise in speech signal. They found that the use of EMD improves the classification performance and efficiency is noticeable.

Sassan Ahmadi and Andreas S Spanias [22] proposed a "cepstrum-based pitch detection using a new statistical voiced/unvoiced classification". In his method voicing decision are made using multi feature voiced unvoiced classification based on statistical analysis of cepstral peak, zero crossing rate and energy of short time segment of speech signal. He found that the performance was improved under noisy conditions.

III. METHODOLOGIES

A. Voiced and Unvoiced Classification Using Zero Crossing Rate and Energy

Combined zero crossing rate and energy calculation zero crossing rate is an important parameter for voiced/unvoiced classification .It is also often used as a part of the front end processing in Automatic speech recognition system. The zero crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced because of excitation of vocal tract by the periodic flow of air at the glottis and usually shows a low zero-crossing count [23], whereas the unvoiced speech is produced by the constriction of the vocal tract narrow enough to cause turbulent airflow which results in noise and shows high zero-crossing count. Energy of a speech is another parameter for classifying the voiced/unvoiced parts. The voiced part of the speech has high energy because of its periodicity and the unvoiced part of speech has low energy.

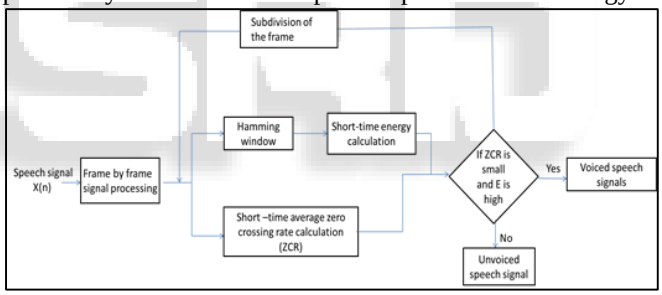


Fig. 1: Block diagram of the voiced/unvoiced classification

The analysis for classifying the voiced/unvoiced parts of speech has been illustrated in the block diagram in Fig.1

B. Voiced and Unvoiced Classification using adaptive thresholding with bivariate EMD

In EMD method, combine speech signal and fractional Gaussian noise (fGn) to form the complex signal. Fractional Gaussian noise (fGn) is a generalization of ordinary discrete white Gaussian noise, and it is a versatile model for broad-band noise. After that apply bivariate EMD .In bivariate EMD, Breaking down signal into various component. It is a decomposing method. The EMD method does not require any condition about the stationary and linearity of the signal .it is suitable for analysis of nonlinear and non-stationary signals. it decomposes the nonlinear and non-stationary signals into set of band limited components known as intrinsic mode function(IMF).

Two conditions of the IMF are as follows:-

1) The number of extreme and the number of zero crossing must either be equal or differ at most by one.

2) The mean value of envelops defined by connecting local maxima and local minima is zero.

After that calculate the log energies of IMF. Conditions for voiced and unvoiced signal.

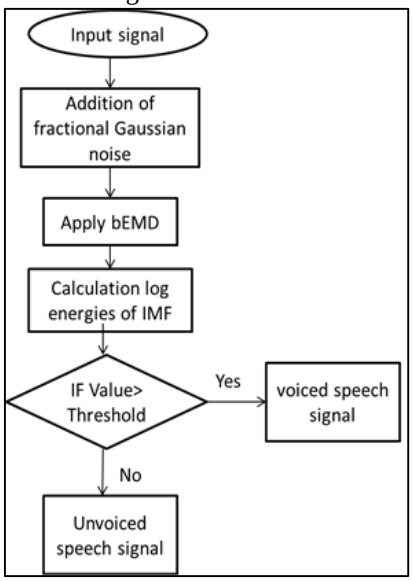


Fig. 2: Flow chart of EMD

Voiced signal-The log energy of at least one IMF will exceed the threshold. if the dominant frequency of that IMF is within the pitch range(50-500Hz) then speech signal is classified as voiced.

Unvoiced signal:-No IMF is expected to have log energy that would exceed the threshold. Rather the log energy distribution is close to that of the fGn.

If sub band log energies IMF will exceed the threshold level then section is marked as voiced otherwise unvoiced.

C. Voiced and Unvoiced Classification using Fuzzy Logic

A method to classify the speech into silence voiced and unvoiced detection using short time energy and zero-crossing in a fuzzy logic system. Figure 1 presents the fuzzy logic system, where the zero-crossing (ZC) and short term energy (STE) are the inputs of fuzzy logic control and the (Detect) is the output. The signal was segmented into frames with duration 10 ms. then; hamming window was applied to prevent discontinuity. The mean of zero-crossing and short term energy was computed for each frame and set as inputs to fuzzy logic control. Voice, unvoiced and silence detection is an output of fuzzy logic control.

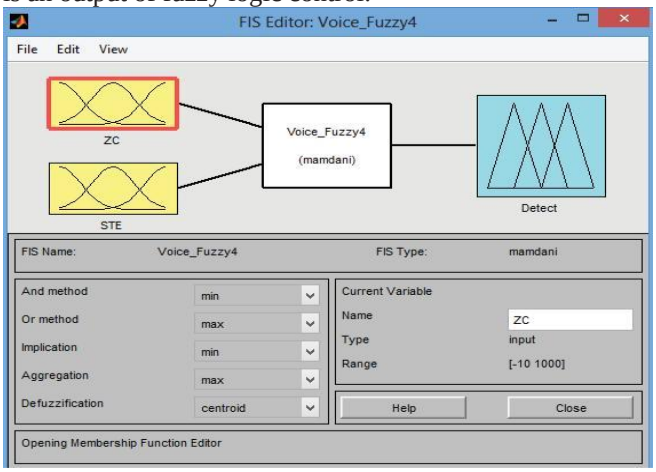


Fig. 3: Fuzzy Logic System

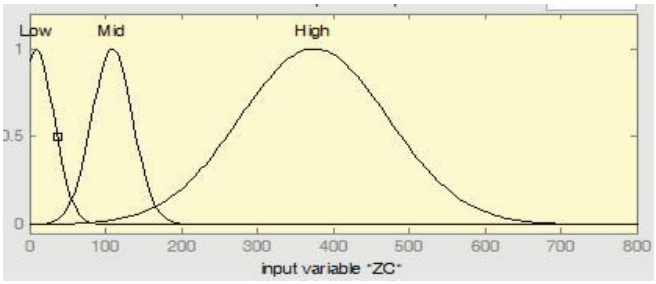


Fig. 4: Membership Function of ZC

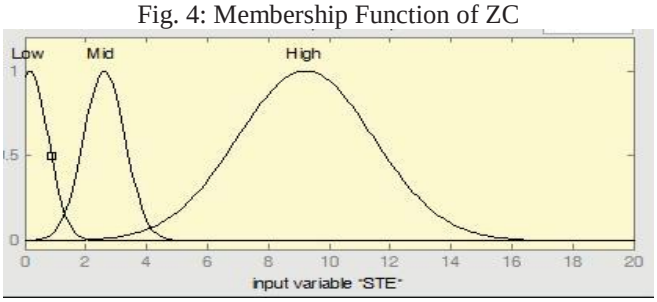


Fig. 5: Membership Function of STE

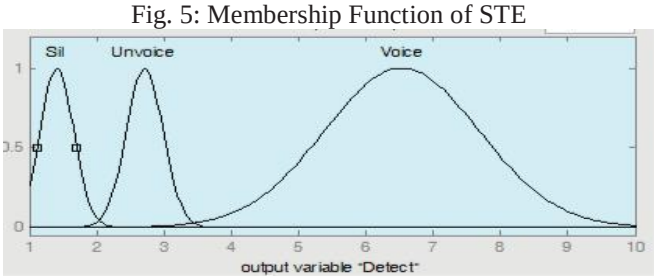


Fig. 6: Membership function of detect

To define the membership function of each linguistic variable we used three membership functions for inputs (ZC and STE). The notation for Zero-crossing is Low, Mid and high as shown in Figure 4. The notation of short term energy (STE) id Low, Mid and High as shown in Figure 5. The notation of fuzzy output (Detect) is Sil, Unvoiced and Voice as shown in Figure 6. The membership functions were tuned after many experiment manually to achieve good results.

D. Voiced and Unvoiced Classification using clustering

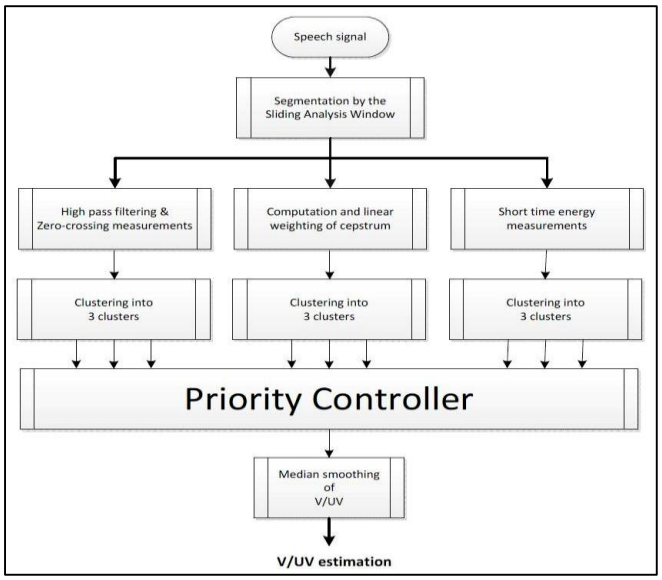


Fig. 7 : Voice/unvoiced classification using clustering

a new method for making v/uv decision is developed which uses a multi-feature v/uv classification algorithm based on the analysis of cepstral peak, zero

crossing rate, and autocorrelation function (ACF) peak of short-time segments of the speech signal by using some clustering methods. This v/uv classifier achieved excellent results for identification of voiced and unvoiced segments of speech.

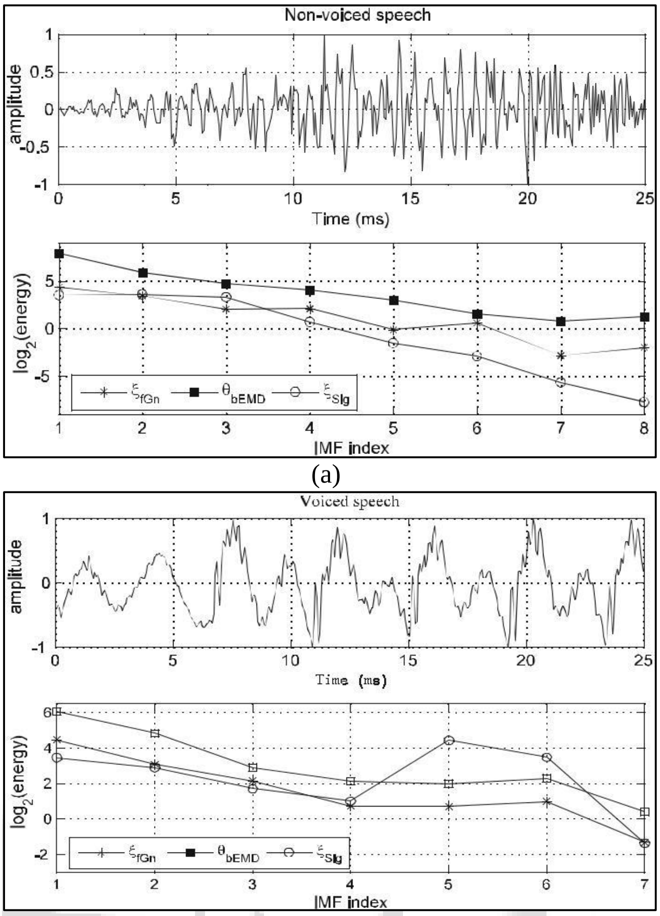
IV. DISCUSSION ON RESULT OBTAINED

A. Voiced and Unvoiced Classification Using Zero Crossing Rate and Energy

Frames				
For word "four", Sampling frequency fs=8000Hz		ZCR	Energy (J)	Decision
Frame-1 (400 Samples)		152	0.0018	Unvoiced
Frame-2	Frame-2 1(200 Samples)	52	0.0543	Unvoiced
	Frame-22(200 Samples)	19	21.1189	Voiced
Frame-3 (400 Samples)		41	186.6628	Voiced
Frame-4 (400 Samples)		41	230.5772	Voiced
Frame-5 (400 Samples)		43	252.98	Voiced
Frame -6(400 Samples)		56	193.70	Voiced
Frame-7	Frame-71(200 Samples)	31	27.2842	Voiced
	Frame-72(200 Samples)	30	25.960	Voiced
	Frame-811(100 Samples)	24	3.4214	Voiced
Frame-8	Frame-812(100 Samples)	11	0.4765	Unvoiced
	Frame-82(200 Samples)	19	0.166	Unvoiced
Frame-9 (400 Samples)		89	0.0054	Unvoiced

Table 1: includes the voiced/unvoiced decisions for word "four." It has 3600 samples with 8000Hz sampling rate

B. Voiced and Unvoiced Classification using adaptive thresholding with Bivariate EMD



(a)

(b)

Fig. 8: Voiced and unvoiced speech discrimination. (a) Voiced speech segment (upper panel), log-energies of IMF5 and IMF6 cross the threshold (lower panel) and it is classified as voiced, (b) non-voiced segment (upper panel), no IMF is found with log-energy greater than the threshold level (lower panel) and hence is it a non-voiced one

C. Voiced and Unvoiced Classification using Fuzzy Logic

In this method for voice and unvoiced detection using fuzzy logic, we used the error difference between voice detected by human and voice detected using this method. The error was calculated using equation (1).

$$error = \frac{|TH-TC|}{TH} \times 100 \quad (1)$$

Where TH is the manually voice frame length detected by human as shown in table 1, and TC is the length of voice detected using the proposed approach.

	Voice frame length detected (ms)												total
TH	0.262	0.26	0.416	0.43	0.19	0.21	0.397	0.17	0.097	0.416	0.445	0.33	3.623
TC	0.3	0.2	0.4	0.4	0.2	0.2	0.4	0.2	0.1	0.4	0.4	0.3	3.5
TH-TC	-0.038	0.06	0.016	0.03	-0.01	0.01	-0.003	-0.03	-0.003	0.016	0.045	0.03	0.093
Error	14.50	23.08	3.85	6.98	5.26	4.76	0.76	17.65	3.09	3.85	10.11	9.09	**2.57**

Table 2: Automatic and manual voice segmentation length of digit file

So from the results in table 1, we can calculate the overall detection error using equation in (2).

$$Overall\ Error = \frac{\sum_{i=1}^{n} |TH-TC|}{\sum_{i=1}^{n} TH} \times 100 \quad (2)$$

Where n is a number of voice frames detected. So from the table 1 above the overall error segmentation is approximate 2.5%.

D. Voiced and Unvoiced Classification using clustering

In this method [18] 821 frames of speech, that were taken from TIMIT, were tested. To calculate the error probability of each of the rules (the six rules described above with considering the priorities we defined), we counted the number of frames that were classified as voiced or unvoiced in each rule (each priority) based on the priorities we determined. Then we counted the number of frames, which were wrongly classified. The frames were labeled visually by looking at their time domain shape and their frequency domain spectrum. The results are depicted in Table 2. Totally the error for voiced segments was 4.8% and the error for unvoiced segments was 1.1%

	auto-1 (5th priority)	ceps-1 (6th priority)	etc
The number of frames identified V or UV	142	69	34
The number of frames wrongly identified	7	10	3

Table 2: Simulations Result [18]

From table 2 it is clearly evident that the number of frames wrongly identified are going ups as priorities further increases. Thus the accuracy of the clustering algorithm cannot be said up to te mark.

V. Conclusion

Various works have been done in the field of voiced/unvoiced classification. The method available for the voiced/unvoiced classifications are based on empirical mode decomposition, wavelet, cepstrum etc and statistical model such as neural network, hidden Markov model (HMM) and Gaussian mixture model (GMM). In all these methods the voiced/unvoiced classification is usually perform by setting a threshold value with some acoustic features say, short time energy, zero crossing rate etc. The main problem is the determination of effective threshold which effect the classification problems performance is also decides the choice of threshold. After throw review and study of existing methods it has been observed that a mixed approach for the voiced/unvoiced classification can improve the performance of the existing schemes. Thus we can conclude that a Hybrid method is required for effective voiced/unvoiced classification.

References

[1] B. Yin, E. Ambikairajah, F. Chen, Voiced/unvoiced pattern-based duration modeling for language identification, in: IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 2009, pp. 4341–4344.
[2] E. Paksoy, J. Carlos de Martin, A. McCree, C.G. Gerlach, A. Anandakumar, W.M. Lai, V. Viswanathan, An adaptive multi-rate speech coder for digital cellular telephony, in: IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Phoenix, USA, vol. 1, 1999, pp. 193–196.
[3] A.M. Kondoz, Digital Speech: Coding for Low Bit Rate Communication Systems, Wiley, England, 2004.
[4] P. Sircar, R.K. Saini, Parametric modeling of speech by complex AM and FM signals, Digital Signal Processing 17 (6) (2007) 1055–1064.
[5] B. Resch, M. Nilsson, A. Ekman, W.B. Kleijn, Estimation of the instantaneous pitch of speech, IEEE Transactions on Audio, Speech and Language Processing 15 (3) (2007) 813–822.
[6] D. Joho, M. Bennewitz, S. Behnke, Pitch estimation using models of voiced speech on three levels, in: IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Honolulu, USA, vol. 4, 2007, pp. 1077–1080.
[7] P.A. Naylor, A. Kounoudes, J. Gudnason, M. Brookes, Estimation of glottal closure instants in voiced speech using the DYPSA algorithm, IEEE Transactions on Audio, Speech and Language Processing 15 (1) (2007) 34–43.
[8] Arifianto D (2007) Dual parameters for voiced–unvoiced speech signal determination. IEEE ICASSP, May 2007, pp 749–752
[9] Molla, M. K. I., Hirose, K., &Hasan,(2015) M. K. Voiced/non-voiced speech classification using adaptive thresholding with bivariate EMD. Pattern Analysis and Applications, 1-6.
[10] Upadhyay, A., &Pachori, R. B. (2015).Instantaneous voiced/non-voiced detection in speech signals based on variational mode decomposition.Journal of the Franklin Institute.
[11] Abrol, V., Sharma, P., & Sao, A. K. (2015). Voiced/nonvoiced detection in compressively sensed speech signals. Speech Communication, 72, 194-207.
[12] Algabri, M., Alsulaiman, M., Muhammad, G., Zakariah, M., Bencherif, M., & Ali, Z. (2015). Voice and Unvoiced Classification Using Fuzzy Logic.In Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV) (p. 416). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
[13] Molla, M. K. I., Hirose, K., Roy, S. K., & Ahmad, S. (2011,). Adaptive thresholding approach for robust voiced/unvoiced classification. In Circuits and Systems (ISCAS), 2011 IEEE International Symposium on (pp. 2409-2412). IEEE.
[14] Jain, P., &Pachori, R. B. (2013). Marginal energy density over the low frequency range as a feature for voiced/non-voiced detection in noisy speech signals. Journal of the Franklin Institute, 350(4), 698-716.
[15] Sharma, P., &Rajpoot, A. K. (2013). Automatic identification of silence, unvoiced and voiced chunks in speech.Journal of Computer Science & Information Technology (CS & IT), 3(5), 87-96.
[16] Senturk, Z., Yetgin, O. E., &Salor, O. (2014). Voiced-unvoiced classification of speech using autocorrelation matrix. In Signal Processing and Communications Applications Conference (SIU), 2014 22nd (pp. 1802-1805). IEEE.

[17] Faycal, Y., &Bensebti, M. (2014). Comparative performance study of several features for voiced/non-voiced classification.Int. Arab J. Inf. Technol., 11(3), 293-299.
[18] Ahmadi, S., &Spanias, A. S. (1999). Cepstrum-based pitch detection using a new statistical V/UV classification algorithm.Speech and Audio Processing, IEEE Transactions on, 7(3), 333-338.
[19] Molla, M. K. I., Hirose, K., &Minematsu, N. (2009). Robust voiced/unvoiced speech classification using empirical mode decomposition and periodic correlation model. In INTERSPEECH (pp. 2530-2533)
[20] Verteletskaya, E., Sakhnov, K., &Šimák, B. (2009). Pitch detection algorithms and voiced/unvoiced classification for noisy speech. In Systems, Signals and Image Processing, 2009.IWSSIP 2009. 16th International Conference on (pp. 1-5). IEEE.
[21] Dhananjaya, N., &Yegnanarayana, B. (2010). Voiced/nonvoiced detection based on robustness of voiced epochs. Signal Processing Letters, IEEE, 17(3), 273-276.
[22] Radmard, M., Hadavi, M., &Nayebi, M. M. (2011). A new method of voiced/unvoiced classification based on clustering. Journal of Signal and Information Processing, 2(04), 336.
[23] Jaber Marvan, "Voice Activity detection Method and Apparatus for voiced/unvoiced decision and Pitch Estimation in a Noisy speech feature extraction", 08/23/2007, United States Patent 20070198251.

