



A Method for Voiced/Unvoiced Classification of Noisy Speech by Analyzing Time-Domain Features of Spectrogram Image

Kazi Mahmudul Hassan^{1,*}, Ekramul Hamid², Khademul Islam Molla²

¹Department of Computer Science & Engineering, Jatiya Kabi Kazi Nazrul Islam University, Mymensingh, Bangladesh

²Department of Computer Science & Engineering, University of Rajshahi, Rajshahi, Bangladesh

Email address:

munnakazi92@gmail.com (K. M. Hassan), ekram_hamid@yahoo.com (E. Hamid), khademul.ru@gmail.com (K. I. Molla)

*Corresponding author

To cite this article:

Kazi Mahmudul Hassan, Ekramul Hamid, Khademul Islam Molla. A Method for Voiced/Unvoiced Classification of Noisy Speech by Analyzing Time-Domain Features of Spectrogram Image. *Science Journal of Circuits, Systems and Signal Processing*.

Vol. 6, No. 2, 2017, pp. 11-17. doi: 10.11648/j.cssp.20170602.12

Received: September 11, 2017; **Accepted:** September 21, 2017; **Published:** October 23, 2017

Abstract: This paper presents a voiced/unvoiced classification algorithm of the noisy speech signal by analyzing two acoustic features of the speech signal. Short-time energy and short-time zero-crossing rates are one of the most distinguishable time domain features of a speech signal to classify its voiced activity into voiced/unvoiced segment. A new idea is developed where frame by frame processing has done in narrow band speech signal using spectrogram image. Two time domain features, short-time energy (STE) and short-time zero-crossing rate (ZCR) are used to classify its voiced/unvoiced parts. In the first stage, each frame of the analyzing spectrogram is divided into three separate sub bands and examines their short-time energy ratio pattern. Then an energy ratio pattern matching look up table is used to classify the voicing activity. However, this method successfully classifies patterns 1 through 4 but fails in the rest of the patterns in the look up table. Therefore, the rest of the patterns are confirmed in the second stage where frame wise short-time average zero-crossing rate is compared with a threshold value. In this study, the threshold value is calculated from the short-time average zero-crossing rate of White Gaussian Noise (wGn). The accuracy of the proposed method is evaluated using both male and female speech waveforms under different signal-to-noise ratios (SNRs). Experimental results show that the proposed method achieves better accuracy than the conventional methods in the literature.

Keywords: Voiced/Unvoiced Classification, Spectrogram Image, Short-Time Energy Ratio, Energy Ratio Pattern, Short-Time Zero-Crossing Rate, White Gaussian Noise

1. Introduction

Speech processing is an interesting area of signal processing where Voiced/Unvoiced classification is one of the classic problems. Considerable efforts have been spent by the researchers in recent years, but results are still not quite satisfactory in case of noisy environments. Speech has several fundamental characteristics in both time-domain and frequency-domain. In Time-domain, speech signal features are short-time energy, short-time zero-crossing rate, and short-time autocorrelation. Speech can be divided into several voiced and unvoiced regions. Short-time energy and short-time zero-crossing rate are most important features to

detect voiced and unvoiced speech in both noisy and noiseless environment. Numerous speech processing applications like speech synthesis, speech enhancement, and speech recognitions are highly dependent on the successful segmentation of speech signal into voiced, unvoiced region.

The human voice frequency is specifically a part of human sound production where the vocal folds (vocal cords) are the primary sound source. More or less constant frequency tones of some duration are consisted in voiced speech, which is made when vowels are spoken by human. Voiced speech is produced when periodic pulses of air generated by the

vibrating glottis resonate through the vocal tract. Approximately two-thirds of speech is voiced which has important intelligibility property. Unvoiced speech is caused by air passing through a narrow constriction of the vocal tract as when consonants are spoken, which is non-periodic, random-like sounds. Because of the periodic nature of voiced speech, it can be identified, and extracted more precisely than unvoiced speech [1].

In recent years considerable efforts have been spent by researchers in solving the problem of classifying speech into voiced/unvoiced segment [2-8]. A pattern recognition approach had been applied to decide whether a given segment of a speech signal should be classified as voiced, unvoiced, or silence, based on measurements of time-domain features of speech [2]. Minimum distance rule was used there to determine the particular class of a speech segment where measured parameters were distributed according to the multidimensional Gaussian probability density function. The major limitation of this method was, it had needed to train the algorithm on every specific set of measurements those were chosen to classify, and for the particular recording condition. It also needs to adapt the means and covariance matrices continuously for better performance in nonstationary speaking environments. A multifeature voiced/unvoiced classification algorithm based on statistical analysis of cepstral peak, zero-crossing rate, and energy of short-time segments of the speech signal was proposed by S. Ahmadi and A. S. Spanias [3]. A binary V/UV classification had proposed which was performed based on three features that can be divided into two categories, features which provide a preliminary V/UV discrimination and a feature which directly corresponds to the periodicity in the input speech.

Y. Qi and Bobby R. Hunt in [4] classified voiced and unvoiced speech using non-parametric methods made by a multilayer feed-forward network which was evaluated and compared to a maximum-likelihood (ML) classifier. However, the network training may take much longer than the calculation of means and covariance matrices for the ML classifier which eventually increasing the computational complexity of the method. L. Siegel et al. in [5] proposed a classifier which viewed voiced/unvoiced classification as a pattern recognition problem where a number of features were used to make that classification. To develop a classifier, a set of features and speaker were used to train the system in a nonparametric, nonstatistical way. Accuracy was significantly biased by the selection of features and speakers in the training set which were needed to be identified and enhanced for the better result, major drawbacks of the method. Childers et al. in [7] proposed an algorithm that was capable of classifying speech into four categories using two-channel (speech and electroglottogram) signal analysis. Here level-crossing rate (LCR) and energy of the EGG signal play role as features to classify them. Various threshold values were determined empirically to distinguish them, which becomes a major drawback of this method. Jashmin K. Shah et al. in [8] had presented two noble approaches to classify voiced/unvoiced speech based on acoustical features and

pattern recognition. The first method is based on Mel frequency cepstral coefficient with Gaussian mixture model (GMM) classifier, and the other was based on Linear Prediction Coefficient (LPC) coefficient and reduced dimensional LPC residual with GMM classifier. The method suffers from false detection problem, which could have occurred if there are less than few pitch periods within a frame in duration.

The proposed method is an attempt to classify noisy speech signal into voiced/unvoiced segment in two stages, using short-time energy ratio pattern and short-time zero-crossing rate. In this paper, we have proposed an approach for speech classification using short-time sub band energy features of spectrogram images of speech signals. In the first stage, the analyzing spectrogram is divided into three separate sub bands (high, low and mid) and calculate their short-time energy ratio. Then an energy ratio pattern matching look up table is used to classify the voicing activity. However, this stage successfully classifies patterns 1 through 4 in the look up table. The remaining patterns of the look up table are confirmed in the second stage where we calculate frame wise short-time average zero-crossing rate (ZCR). In the method, the short-time average zero-crossing rate of White Gaussian Noise (wGn) is used to estimate a threshold value which is compared with the calculated short-time ZCR of the speech signal. So this stage confirms the voicing decision if the first stage fails. The analysis and methods that are used in this study have presented in the second and third part. The experimental conditions & results are given in the fourth part.

2. Analysis of Speech Signal for Voiced/Unvoiced Classification

Analyzing speech signal based on energy, it is assumed that a voiced segment of speech will have higher short-time energy and will have lower when an unvoiced segment occurs. Also in the energy distribution of spectrogram image, the voiced segment has clear harmonics whereas they abruptly change in unvoiced segments.

Typical adult males have a fundamental frequency from 85 to 180 Hz and that of a typical adult female from 165 to 255 Hz in voiced speech [12], [13]. Those fundamental frequencies contain enough of the harmonic series, which contains most of the energy components and in most of the cases, this energy gradually degraded from lower frequency band to upper. Besides unvoiced sound contains a low amount of energy compare to voiced one and it shows no specific pattern, rather contains most energy in mid and higher frequency band with abrupt changes and higher amount of zero-crossing rate. In case of noisy speech signals, most of the unvoiced components mixed with the noise which makes it more difficult to segregate and this difficulty level increase with the strength of noise in the speech signal.

An important parameter, short-time Zero-Crossing Rate, is an indicator of the signal spectrum frequency in which the

energy is concentrated. Zero-Crossing is counted on discrete time signal if successive samples have different algebraic signs [10]. Zero-crossing rate is a measure of the number of times amplitude of a speech signal passes through a value of zero within a given time interval/frame is shown in Figure 1.

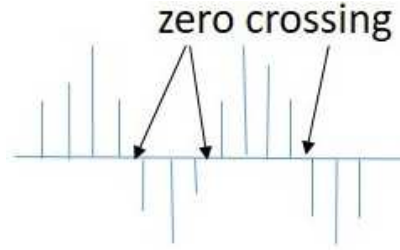


Figure 1. Definition of Zero-crossing rate [10].

Zero-crossing rate can be defined as [10]:

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m) \quad (1)$$

Where

$$sgn[x(n)] = \begin{cases} -1, & x(n) < 0 \\ 1, & x(n) \geq 0 \end{cases}$$

$$\text{and } w(n) = \begin{cases} \frac{1}{2W}, & 0 \leq n \leq W-1 \\ 0, & \text{otherwise} \end{cases}$$

W = number of samples within a frame

Previous studies show an approximate distribution of the

short-time zero-crossing rate in voice/unvoiced segments [10]. In case of clean speech samples, the zero-crossing rate is very low in silent region, low in voiced and generally high in unvoiced region [9], [10].

Here in this study, a zero-crossing rate (ZCR) experiment has taken on white Gaussian Noise where it is considered as degrading noise in case of speech signal [11]. In this experiment, 1000k random wGn samples are generated in different strength (dB) and their mean ZCR in per millisecond are calculated using overlapping frames (Frame length 10ms) over those samples. According to that experiment result, mean ZCR of wGn is approximately 3.8 per ms, and it does not depend or vary on the strength of wGn. In noisy speech sample, we can assume that voiced components ZCR will have less than 3.8 per ms in most of the cases with particular energy changing pattern.

3. Proposed Method for Voiced/Unvoiced Classification

In our proposed design, we combined short-time energy and short-time zero-crossing rate of a spectrogram image. The analysis for classifying the voiced/unvoiced parts of speech is illustrated in the block diagram Figure 2.

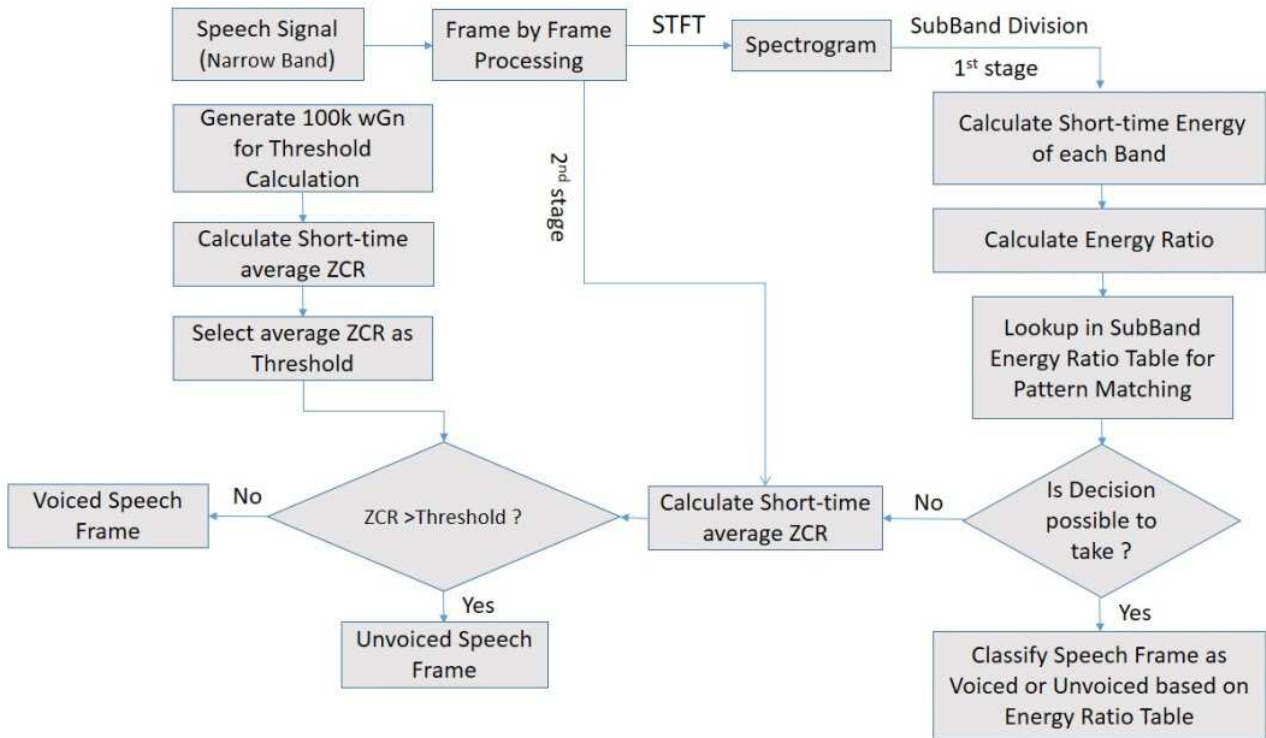


Figure 2. Block Diagram of the Proposed Method.

In our proposed method, a narrow band (4 kHz max) speech signal is processed frame by frame (15 ms frame size with 50% overlapping between consecutive frames). To

represent the instantaneous spectrum of a signal $\{X\}$, we use Short-Time Fourier Transform (STFT) which results to the spectrogram that shows the evolution of frequencies in time.

It is observed that small value for time steps returns a large spectrogram, which requires a long computation time. The time resolution and the frequency resolution of the STFT depend on window length.

The Speech production model suggests because of the spectrum fall-off introduced by the glottal wave, the energy of voiced speech is concentrated below about 3 kHz, whereas most of the energy is found at higher frequencies for unvoiced speech [10]. Based on those findings, each frame of the analysis spectrogram is divided into three separated sub band using two thresholds $th1=1200$ Hz and $th2=3000$ Hz which are selected experimentally. Here the band limit of sub bands are shown below:

Table 1. Frequency Range of Each Sub Band.

Band Name	Start Freq. (Hz)	End Freq. (Hz)
Low Band	0	1200
Mid Band	1201	3000
High Band	3001	4000

In the next step, the cumulative energy of each band EL (Low band), EM (Mid Band) and EH (High Band) is calculated for each frame (see Figure 3). Then we examine the energy ratio of the three sub bands with the Energy Ratio Pattern table and take a decision whether the frame is a voiced or unvoiced. If it fails to decide, we need to calculate the short-time ZCR of that frame to make decisions further.

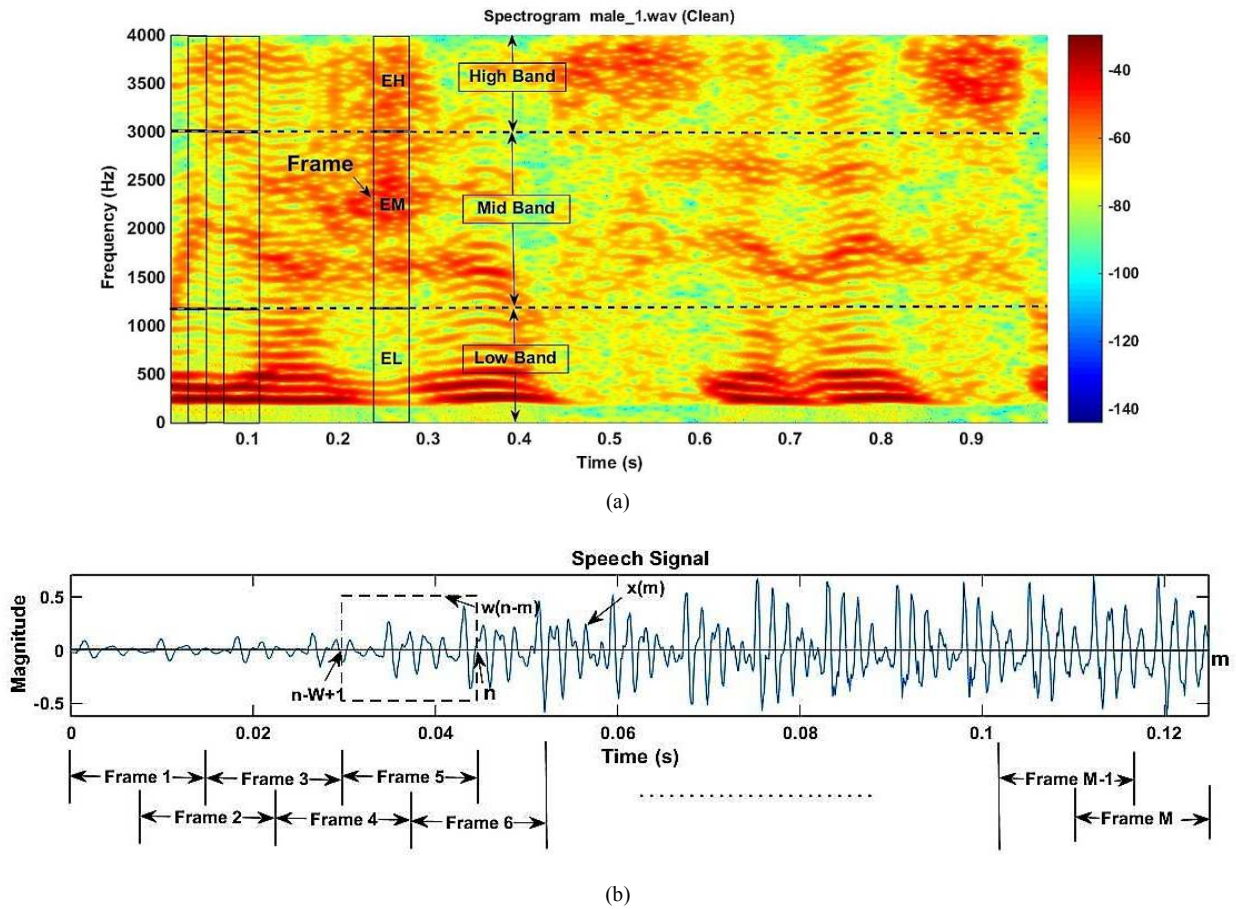


Figure 3. a. Spectrogram with Frame & Sub Band Energy. b. Speech Signal with Window and Overlapping Frames.

In general, using the following formulas, we can calculate the short-time energy of three sub bands EL, EM, and EH from the spectrogram, where EL_f , EM_f , and EH_f represents sub band energy of f^{th} frame respectively.

$$EL_f = \sum_{b=1}^{b=TH1} P_{(f,b)} \quad (2)$$

$$EM_f = \sum_{b=TH1+1}^{b=TH2} P_{(f,b)} \quad (3)$$

$$EH_f = \sum_{b=TH2+1}^{b=N} P_{(f,b)} \quad (4)$$

Where

f = Frame index, $f=1, 2, 3, \dots, M$

M = Number of frames

$P(f, b)$ = Power Spectral Density of bin b , frame f

N = Number of frequency bins in Spectrogram.

$TH1$ = Threshold bin for Low band, which is proportional to threshold $th1$ and depends on N .

$TH2$ = Threshold bin for Mid band, which is proportional to threshold $th2$ and depends on N .

Now if $EL_f = 5.02 \times 10^{-4}$, $EM_f = 3.55 \times 10^{-4}$, and $EH_f = 2.95 \times 10^{-4}$ then we can represent them for f^{th} frame as

$$EL_f: EM_f: EH_f = 5.02: 3.55: 2.95 \quad (5)$$

Where $EL > EM > EH$ is the Energy Ratio Pattern of f^{th} frame.

Table 2. Energy Ratio Pattern Table.

Pattern Type	Energy Ratio Pattern	Decision	ZCR Calculation
1	$EH < EM < EL$	Voiced	No need
2	$EH > EM > EL$	Unvoiced	No need
3	$EH > EL \&\& EL > EM$	Unvoiced	No need
4	$EM > EH \&\& EH > EL$	Unvoiced	No need
5	$EM > EL \&\& EL > EH$	Not sure	Yes
6	$EL > EH \&\& EH > EM$	Not sure	Yes
7	Others	Not sure	Yes

From the Table 2, if the sub band energy ratio satisfies any of the first four conditions, we can take the decision without calculating the ZCR. But if it is not, then we need to calculate ZCR of that frame to take the decision whether it is voiced or unvoiced. In the experiment, the ZCR threshold is considered 3.8 per ms, used in the situations where the decision cannot make concretely. In such case, the frame is classified as voiced, if its short-time ZCR rate is below the threshold. Otherwise, it is considered as unvoiced frame.

4. Experiment Details & Results

The experiments are performed on 3 male and 3 female clean speech from TIMIT database, where each sample has 1sec duration and sampling frequency is 8 kHz. White

Gaussian noise is used as a degrading source and added to the signal in different SNRs. The frame duration is 15 milliseconds and 50% overlapped between consecutive frames. So each frame contains 120 sample points and 60 sample points are overlapped. In case of reconstruction, if the consecutive frames are voiced-unvoiced or reverse one, the overlapping samples have to be divided into two parts (30 sample points each) and add to the respective frames. In the experiment, the speech sample contains only voiced and unvoiced speech without silence. The algorithm is developed and tested for narrow band. So in case of wide band speech signal, that signal has to be passed throw bandpass filter to generate narrow band speech signal.

The outcome of the proposed method is given in Table 3 with accuracy in percentage. Here we can see that in most of the cases, clean speech has accuracy more than 96% and this value will increase if we take sample length longer enough. The Average accuracy of the proposed method is approximately 92% which is quite high in case of noisy speech signal compare to other methods [14], [15]. Most of the errors occurred in voiced to unvoiced or unvoiced to voiced transaction frames. If we ignore those transition error where Voiced/Unvoiced characteristics are not concrete, the accuracy of the proposed method will increase further for noisy speech, calculated up to 99.24% for clean speech.

Table 3. Calculated Accuracy of Proposed Method.

Male		Female	
Sample Name: male_1.wav		Sample Name: female_1.wav	
SNR (wGn)	Accuracy (%)	SNR (wGn)	Accuracy (%)
Clean	98.48	Clean	94.70
20 dB	94.70	20 dB	95.45
15 dB	97.73	15 dB	96.21
10 dB	98.48	10 dB	90.15
5 dB	99.24	5 dB	86.36
Sample Name: male_2.wav		Sample Name: female_2.wav	
SNR (wGn)	Accuracy (%)	SNR (wGn)	Accuracy (%)
Clean	94.70	Clean	96.21
20 dB	98.48	20 dB	92.42
15 dB	93.94	15 dB	85.61
10 dB	82.58	10 dB	82.58
5 dB	84.09	5 dB	76.52
Sample Name: male_3.wav		Sample Name: female_3.wav	
SNR (wGn)	Accuracy (%)	SNR (wGn)	Accuracy (%)
Clean	98.48	Clean	98.48
20 dB	97.22	20 dB	96.21
15 dB	96.21	15 dB	93.18
10 dB	90.21	10 dB	94.70
5 dB	87.12	5 dB	86.36

In Figure 4 the performance of the proposed method in both clean and noisy conditions is showed. From both spectrograms, we can see because of wGn, the distribution of energy changes in case of noisy spectrogram but the impact is quite uniform all over the spectrogram. So it does not change the energy ratio pattern, that's why the proposed method able to recognizes voiced/unvoiced segments in noisy environment like clean speech.

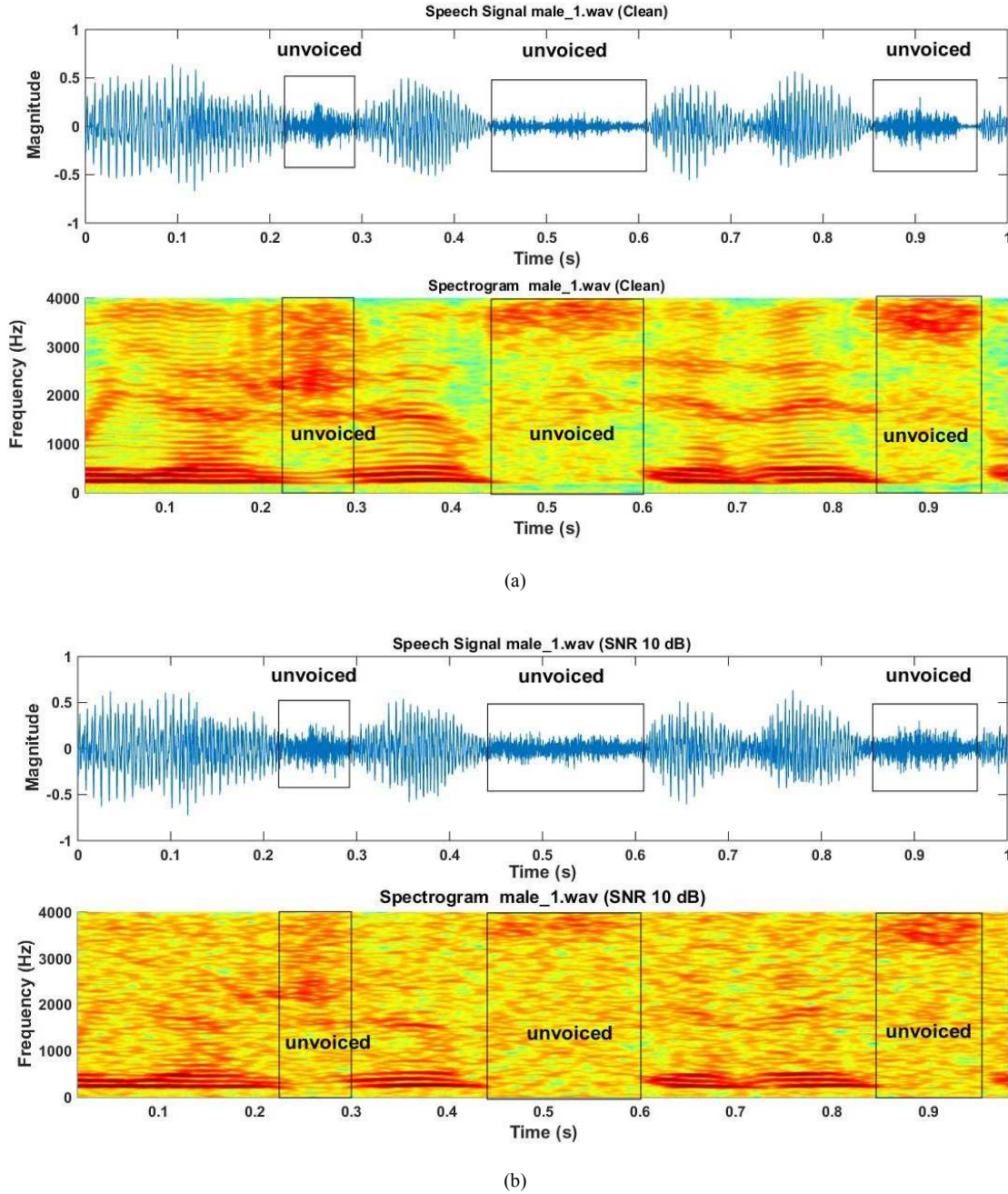


Figure 4. Voiced/Unvoiced Classification Using Proposed Method (a) Clean Speech. (b) SNR 10 dB Speech Sample.

5. Conclusion

To classify a speech signal into voiced/unvoiced segment, a joint approach using STFT, short-time energy and short-time ZCR has presented here. Speech signal was processed frame by frame on spectrogram image, divided it into sub bands and calculated their energy ratio. Classification decision was taken based on their pattern using an energy ratio pattern matching lookup table. We had considered ZCR for confirmed classification in some cases. Further improvements can be made by more study on the sub bands threshold value which had taken here experimentally. A more statistical analysis should have to be done to implement it on

wideband speech signal where new threshold values need to be extracted. Further study on the short-time energy ratio pattern will definitely contribute more on the accuracy level and robustness of the method.

References

- [1] Jong Kwan Lee, Chang D. Yoo, "Wavelet speech enhancement based on voiced/unvoiced decision", Korea Advanced Institute of Science and Technology The 32nd International Congress and Exposition on Noise Control Engineering, Jeju International Convention Center, Seogwipo, Korea, August 25-28, 2003.

- [2] B. Atal, and L. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," IEEE Trans. On ASSP, vol. ASSP-24, pp. 201-212, 1976.
- [3] S. Ahmadi, and A. S. Spanias, "Cepstrum-Based Pitch Detection using a New Statistical V/UV Classification Algorithm," IEEE Trans. Speech Audio Processing, vol. 7 No. 3, pp. 333-338, 1999.
- [4] Y. Qi, and B. R. Hunt, "Voiced-Unvoiced-Silence Classifications of Speech using Hybrid Features and a Network Classifier," IEEE Trans. Speech Audio Processing, vol. 1 No. 2, pp. 250-255, 1993.
- [5] L. Siegel, "A Procedure for using Pattern Classification Techniques to obtain a Voiced/Unvoiced Classifier", IEEE Trans. on ASSP, vol. ASSP-27, pp. 83-88, 1979.
- [6] T. L. Burrows, "Speech Processing with Linear and Neural Network Models", Ph.D. thesis, Cambridge University Engineering Department, U.K., 1996.
- [7] D. G. Childers, M. Hahn, and J. N. Larar, "Silent and Voiced/Unvoiced/Mixed Excitation (Four-Way) Classification of Speech," IEEE Trans. on ASSP, vol. 37 No. 11, pp. 1771-1774, 1989.
- [8] Jashmin K. Shah, Ananth N. Iyer, Brett Y. Smolenski, and Robert E. Yantorno "Robust voiced/unvoiced classification using novel features and Gaussian Mixture model", Speech Processing Lab., ECE Dept., Temple University, 1947 N 12th St., Philadelphia, PA 19122-6077, USA.
- [9] 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.
- [10] Rabiner, L. R., and Schafer, R. W., Digital Processing of Speech Signals, Englewood Cliffs, New Jersey, Prentice Hall, 512-ISBN-13: 9780132136037, 1978.
- [11] Karen Kafadar, "Gaussian white-noise generation for digital signal synthesis" IEEE Transactions on Instrumentation and Measurement, Volume: IM-35, Issue: 4, Dec. 1986 DOI: 10.1109/TIM.1986.6499122
- [12] Titze, I. R. "Principles of Voice Production", Prentice Hall (currently published by NCVS.org) (pp. 188), 1994, ISBN 978-0-13-717893-3.
- [13] Baken, R. J. "Clinical Measurement of Speech and Voice". London: Taylor and Francis Ltd. (pp. 177), 1987, ISBN 1-5659-3869-0.
- [14] Alkulaibi, A., Soraghan, J. J., and Durrani, T. S., "Fast HOS based simultaneous voiced/unvoiced detection and pitch estimation using 3-level binary speech signals", in the proceedings of 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing, pp. 194-197, 1996.
- [15] Lobo, and Loizou, P., "Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition", in the Proceedings of ICASSP, pp. 820-823, 2003.