Short-Time Energy, Magnitude, Zero Crossing Rate and Autocorrelation Measurement for Discriminating Voiced and Unvoiced segments of Speech Signals

Madiha Jalil

School of Engineering,
University of Management and
Technology, Lahore, Pakistan
ms09141@gmail.com

Faran Awais Butt

School of Engineering,
University of Management and
Technology, Lahore, Pakistan
faranawais@gmail.com

Ahmed Malik

School of Engineering,
University of Management and
Technology, Lahore, Pakistan
xs2ahmedmalik@gmail.com

Abstract—**This paper presents different methods of separating voiced and unvoiced segments of a speech signals. These methods are based on short time energy calculation, short time magnitude calculation, and zero crossing rate calculation and on the basis of autocorrelation of different segments of speech signals. From theoretical studies, it has been observed that energy and magnitude for voiced segments is high, whereas ZCR rate is low for voiced signals. Autocorrelation function is used here to show that the voiced segment of speech remains periodic after applying autocorrelation function, while unvoiced signals lose their periodicity. Experimental results have been presented in this paper to verify theoretical studies.**

Keywords—Zero Crossing Rate, Short Time Energy, Autocorrelation, Voiced, Unvoiced

I. INTRODUCTION

Speech is incorporated of several voiced, unvoiced and silence segments. Speech is a non stationary signal but it remains nearly unvaried for small segments i.e. for 10 to 20 ms [8]. This distribution of speech signal in different segments such as voiced , unvoiced and silence gives an elementary acoustic segmentation for many processing applications e.g., speech synthesis, speech recognition and speech enhancement[1]. Voiced speech is consisted of almost constant spectrum. Constant frequencies are made usually when vowels are spoken. The voiced region of speech is generated when vibrating glottis resonates through the vocal tract and produce periodic pulses of air. Periodic pulses are generated at frequencies that are dependent upon vocal tract shape. The major part of speech is voiced. Voiced speech is more important for intelligibility of speech. Non periodic sounds generated by the passage of air that is obstructed by the narrow constriction of the vocal tract when consonants are spoken. So the voiced signals can be distinguished and separated due to its periodicity [2].

This paper presents classification of different speech segments on the basis of energy, magnitude, zero crossing rate and the knowledge that unvoiced segments are non-periodic.
Suggested approaches for discrimination of voiced and unvoiced parts are given in section 3 and the results are given in section 4. The experiment has been performed on two different sounds. One is male's sound and the other is female's. We had a choice between "Hamming" and "Rectangular window", here both windows have been used for achieving results and to make the decision that which one is better. The experimental results also give an insight that how energy, magnitude and other parameters vary by varying the length of the window.

II. TECHNIQUES

A. Short-Time Energy

The short time energy is the energy of short speech segment [5]. Short time energy is a simple and effective classifying parameter for voiced and unvoiced segments [7]. Energy is also used for detecting end points of utterance [10]. The long term definition of signal energy is as below (from reference [3]):

$$E = \sum_{m=-\infty}^{\infty} x^2(m) \qquad (1)$$

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) = x^2(n-N+1) + \ldots + x^2(n) \qquad (2)$$

In above expression E represents energy of the signal x(m). There is a very small or almost no efficacy of this definition for time-varying signals, such as speech.

For a short-term speech signal, an n-th frame window is applied on this signal:

$$x_n(m) = x(m)w(n-m) \qquad n-N+1 \le m \le n \qquad (3)$$

n=0, 1T, 2T,..., N is the window length and T is the frame-shift.

The short time energy of above signal can be determined from following expression:

$$E_n = \sum_{m=n-N+1}^{n} [x(m)w(n-m)]^2 \qquad (4)$$

Where w (n-m) is the window, n is the sample that the analysis window is centred on, and N is the window length. Chosen window selects the interim for processing and slides across progression of squared values. Here high energy would be classified as voiced and lower energy as unvoiced.

B. Short-Time Magnitude

Average magnitude function is defined as (from reference [3])

$$M_n = \sum_{m=-\infty}^{m=\infty} |x(m)|\, w(n-m) \qquad (5)$$

The high energy would be classified as voiced and lower as unvoiced.

C. Zero Crossing Rate

It counts the no of zero crossings in speech signal. Voiced segments have low ZCR as compare to ZCR of unvoiced segments [9].

Consecutive samples in a speech signal have different algebraic signs. ZCR is measurement of "frequency composition" of a signal; this is more valid for narrowband signals such as sinusoids.

The sinusoid of F_O frequency with sampling rate F_S will have F_S / F_O samples per cycle with two zero crossings per cycle. It will result in zero crossing rates.

$$Z = \frac{2F_0}{F_s} \qquad (6)$$

Z is the Crossings/ sample.

Speech signals are broadband signals and the interpretation of Average ZCR is therefore much less precise. However, rough estimates of the spectral properties can be obtained using a representation based on the short time average ZCR. An appropriate definition of computations is

$$\begin{aligned} Z_n &= \sum_{m=-\infty}^{\infty} |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]|\, w(n-m) \\ &where \qquad (7) \\ \operatorname{sgn}[x(n)] &= 1 \qquad x(n) \ge 0 \\ &= -1 \qquad x(n) < 0 \end{aligned}$$

and

$$\begin{aligned} w(n) &= 1/(2N) \qquad 0 \le n \le N-1 \\ &= 0 \qquad otherwise \end{aligned}$$

Several features of speech can be drawn using magnitude, energy and zero crossing. Zero crossing rate is very useful for discriminating speech from noise and for determining start and end of speech segment. Lower energy in zero crossing rate would be classified as voiced and high energy as unvoiced.

D. Short-Time Auto Correlation

Short-time autocorrelation is defined as (from reference [3])

$$R_n(k) = \sum_{m=0}^{m=N-1-k} [x(n+m)w'(m)][x(n+m+k)w'(k+m)] \qquad (8)$$

This expression represents the autocorrelation of a speech segment with specified window length. Rectangular or hamming window can be used, here rectangular window is used. If the window size is reduced the result of auto correlation attenuates. As the number of samples reduces due to shorter window, so this attenuation is quite expected. Voiced segment would be periodic and unvoiced would be non periodic in short time autocorrelation.

III. Simulations Performed For Separating Voiced & Unvoiced segments

For obtaining results, two softwares have been used. One is Matlab 7.10.0.499 and the other is Praat. Matlab is used for programming the algorithms for energy, magnitude, ZCR and autocorrelation. While Praat is used for recording sounds. Praat supports audio files with .wav extension [6].These recorded sounds were read in Matlab, using wavread command. And then various algorithms were applied on these signals to discriminate voice and unvoiced segments of the speech.

The speech signal, presented in Fig.1, is recorded by a female and the sentence is "Hello madiha" and that in Fig.1a is spoken by a male and the sentence is "hello what is your name".

Fig. 2(a, b, c, d, e, f) shows graphical results of short time energy calculation. It can be seen from the figures as we increased the window size the energy waveform becomes smoother while at smaller window, energy fluctuates a lot. So it is clear from the experimental results that suitable choice of window is quite important. Window used for obtaining short time energy is hamming window. Fig 2(a, b and c) represents short-time energy of a female voice and Fig. 2(d, e and f) represents that of male voice. It can also be observed that energy for the voiced segment is higher than that of unvoiced signal. So energy measurement is an efficient way of discriminating voiced and unvoiced segments.

Fig. 3 (all parts) shows results for magnitude calculation. By increasing the size of window, the magnitude for speech signals becomes smoother. Rectangular window is used for measurement of short-time magnitude. It is clear from the results that magnitude for voiced signals is relatively higher than that of unvoiced signals.

Window used for measurement of zero crossing rate is hamming window. Fig. 4 (all parts) shows zero crossing rate for sound signals spoken by male and female.

Fig.5 (all parts) shows autocorrelation results of different segments of a sound signal. Sound signals, presented in Fig 1a and 1b, and were first divided in voiced and unvoiced segments, on the basis of knowledge of their energy, magnitude and zero crossing rate. Then the auto correlation function of the divided segments was taken. From the results, it can be observed that after the application of auto correlation function voiced segment remains periodic , while unvoiced turns out to be non periodic. The figures shown below are the simulations performed on MATLAB for male and female voices. Magnitude energy, zero crossing rate and auto correlations have been simulated in the figures.

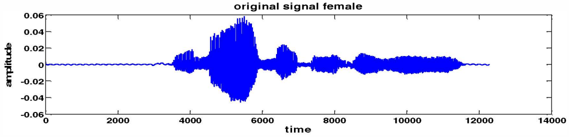

Fig. 1 (a)

Graph Of Sound Recorded ‘Hello Madiha’ in Female voice

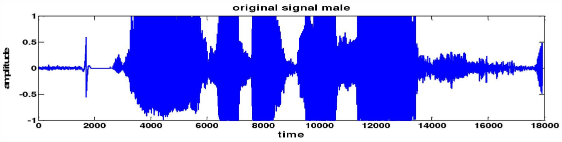

Fig. 1 (b)

Graph Of Sound Recorded ‘Hello what is your name’ in male voice

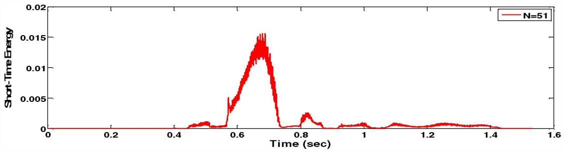

Fig. 2(a)

Short Time Energy Graph of Female voice at window length N=51

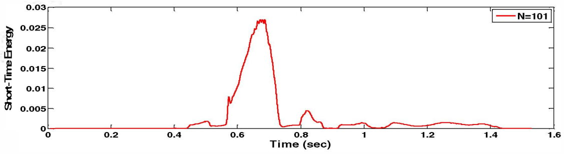

Fig. 2 (b)

Short Time Energy Graph of Female voice at window length N=101

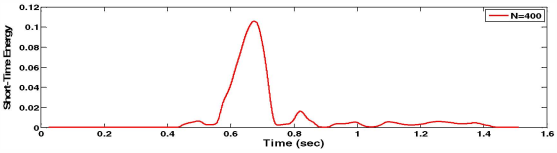

Fig. 2 (c)

Short Time Energy Graph of Female voice at window length N=400

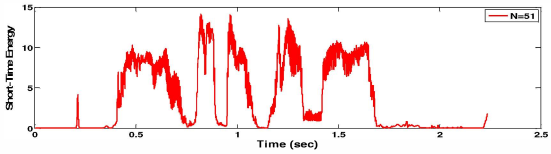

Fig. 2 (d)

Short Time Energy Graph of male voice at window length N=51

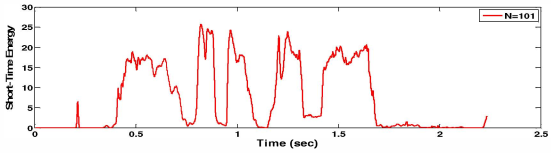

Fig. 2 (e)

Short Time Energy Graph of male voice at window length N=101

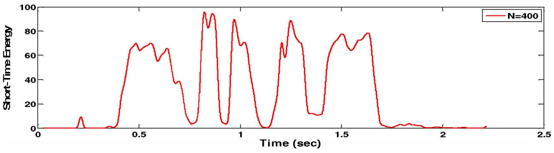

Fig. 2 (f)

Short Time Energy Graph of male voice at window length N=400

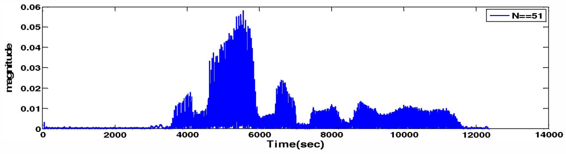

Fig. 3 (a)

Short Time magnitude Graph of Female voice at window length N=51

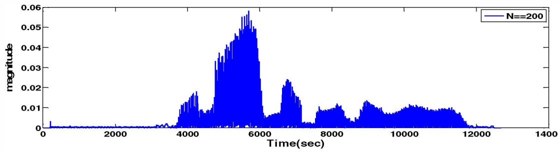

Fig. 3 (b)

Short Time magnitude Graph of Female voice at window length N=200

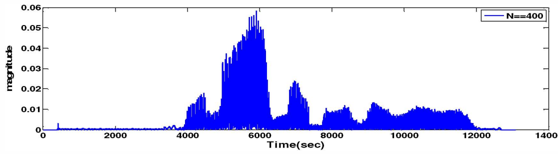

Fig. 3 (c)

Short Time magnitude Graph of Female voice at window length N=400

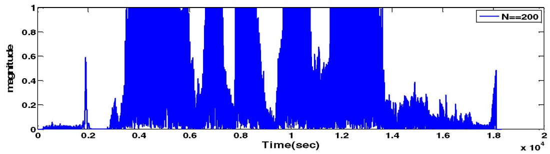

Fig. 3 (d)

Short Time magnitude Graph of male voice at window length N=200

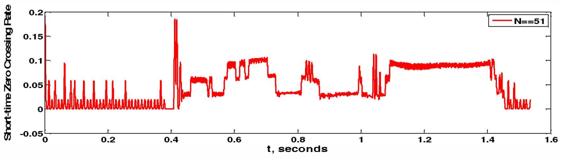

Fig. 4 (a)
Zero Crossing Rate Graph of Female voice at window length N=51

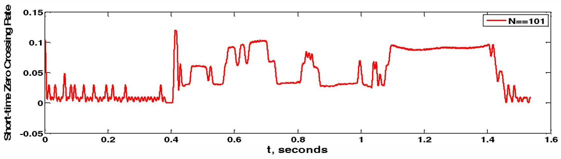

Fig. 4 (b)
Zero Crossing Rate Graph of Female voice at window length N=101

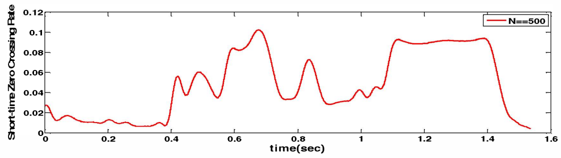

Fig. 4 (c)
Zero Crossing Rate Graph of Female voice at window length N=500

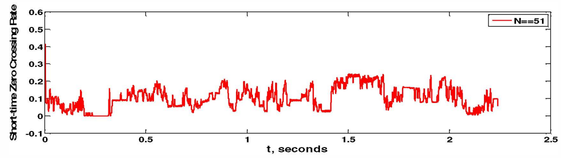

Fig. 4 (d)
Zero Crossing Rate Graph of male voice at window length N=51

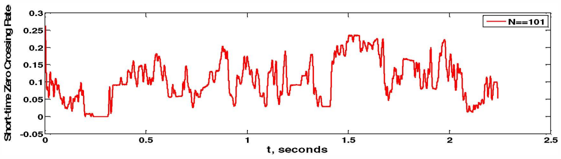

Fig. 4 (e)
Zero Crossing Rate Graph of male voice at window length N=101

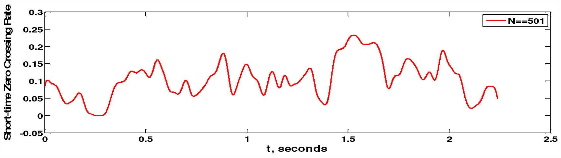

Fig. 4 (f)
Zero Crossing Rate Graph of male voice at window length N=500

The following figures show the zero crossing rates of different segments of male and female voice at different values of N. The auto correlation of voiced and unvoiced signals have been shown in fig.5 (a, b, c, d, e) . From the results it can be observed that energy and magnitude for voiced segments is high as compared to unvoiced segments, whereas ZCR rate is low for voiced signals. Autocorrelation function is used here to show that the voiced segment of speech remains periodic after applying autocorrelation function while unvoiced signals lose their periodicity.

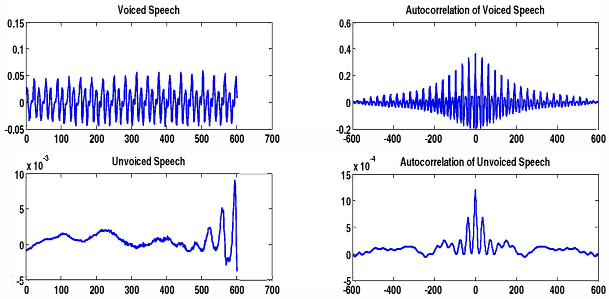

Fig. 5(a)
Zero Crossing Rate Graph of Different Segments of Female voice at window length N=600

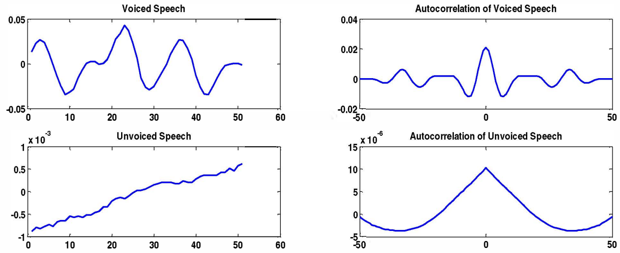

Fig. 5 (b)
Zero Crossing Rate Graph of Different Segments of Female voice at window length N=50

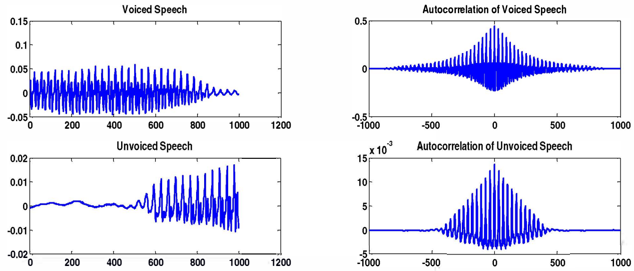

Fig. 5(c)
Zero Crossing Rate Graph of Different Segments of Female voice at window length N=1000

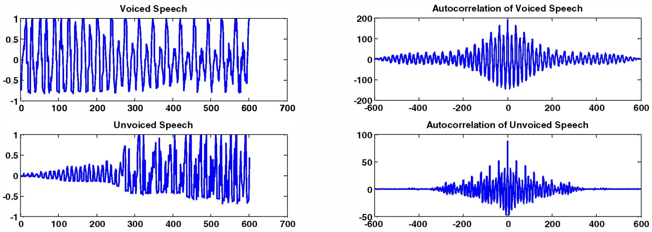

Fig. 5 (d)
Zero Crossing Rate Graph of Different Segments of male voice at window length N=600

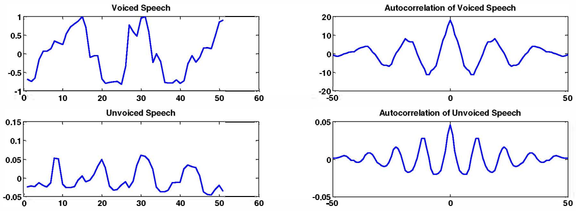

Fig. 5 (e) Zero Crossing Rate Graph of Different Segments of male voice at window length N=50

IV. CONCLUSION

The paper gives experimental verification of the short- time analysis techniques for classification of voice and unvoiced segments of speech signals. The algorithms used are efficient and simple. We used simple techniques of calculating short time energy, Magnitude and Zero crossing rates for classifying voiced and unvoiced segments. We found that a better choice of window and window length gives us more appropriate results. So while separating voiced and unvoiced segments of speech, window and window length should be chosen carefully. This work can be further extended to discriminate silence and noise from the speech signals. Pitch measurement method can also be exploited as a further work.

REFERENCES

[1] R.G. Bachu, S. Kopparthi, B. Adapa, B.D. Barkana, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy," IEEE International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE'08).

[2] J. K. Lee, C.D. Yoo, "Wavelet speech enhancement based on voiced/unvoiced decision," The 32nd International Congress and Exposition on Noise Control Engineering, Jeju International Convention Centre, Seogwipo, Korea, August 25-28, 2003.

[3] Speech Signal Processing, School of Electronic Information, Chapter 3, Wuhan University.

[4] K.Ubul, A.Hamdulla, A.Aysa, "A Digital Signal Processing teaching methodology using Praat," 4th International Conference on Computer Science and Education, 2009.

[5] X.Yang, B.Tan, J.Ding, J.Zhang, J.Gong, "Comparative Study on Voice Activity Detection Algorithm," International Conference on Electrical and Control Engineering, ICECE, 2010.

[6] M.R.Velankar, H.V.Sahasrabuddhe, "Exploring Data Analysis in Music using tool praat," First International Conference of Emerging Trends in Engineering and Technology, 2008.

[7] D.Enqing, L.Guizhong, Z.Yatong, C.Yu, "Voice Activity Detection Based on Short-Time Energy and Noise Spectrum Adaptation," 6th International Conference on Signal Processing, 2002.

[8] D.Enquing, L.Guizhong,_Z.Yatong,_C. Yu, "Voice Activity Detection Based on Short-time Energy and Noise Spectrum Adaptation," 6th International Conference on Signal Processing, August 2002.

[9] W.P.Ng, J.M.H.Elmirghani, R.A.Cryan, C.C.Yoong, S.Broom,"Divergence detection in a speech-excited in-service non-intrusive measurement device," IEEE International Conference on Communications, ICC 2000.

[10] Y.K.Lau, C.K.Chan, "Speech Recognition Based on Zero Crossing Rate and Energy," IEEE Transactions on Acoustic, Speech and Signal Processing, Feb.198.