International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Analysis of Speech Signal Using Graphic User Interface

Solly Joy[1], Savitha Upadhya[2]
EXTC Department, FCRIT, sollyjoy42@gmail.com
EXTC Department, FCRIT, savivashi@gmail.com

Abstract— In this paper, the concepts of speech processing algorithms for speech signal analysis is presented. Speech analysis is performed using short-time analysis to extract features in time domain and frequency domain. The short time domain analysis is useful for computing the time domain features like energy and zero crossing rate. The different frequency or spectral components that are present in the speech signal are not directly apparent in the time domain. Hence the frequency domain representation using Fourier representation is needed. The time varying nature of spectral information in speech leads to the need for short time of Fourier transform, termed more commonly as Short time Fourier Transform (STFT).The effect of different types of windows used in short time analysis with and without overlapping and the effect of window length in speech analysis are also demonstrated.

Keywords- Short time analysis, Windowing, Short time energy, Short time magnitude, Short time zero crossing rate, Short time autocorrelation.

I. INTRODUCTION

Speech processing applications uses certain features of speech signals in accomplishing their tasks. The extraction of these features and to obtain them from a speech signal is known as speech analysis. It can be done in time domain as well as frequency domain. Analyzing speech in the time domain often requires simple calculation and interpretation. The frequency domain provides the mechanisms to obtain the most useful parameters in speech analysis. Most models of speech production assume a noisy or periodic waveform exciting a vocal-tract filter. The excitation and filter can be described in either the time or frequency domain, but they are often more consistently and easily handled spectrally. Voiced speech consists of periodic or quasi periodic sounds made when there is a significant glottal activity. Unvoiced speech is non periodic, random excitation sounds caused by air passing through a narrow constriction of the vocal tract. Unvoiced sounds include the main classes of consonants which are voiceless fricatives and stops. When both quasi-periodic and random excitations are present simultaneously, the speech is classified voiced because the vibration of vocal folds is part of the speech act [1]. In other contexts, the mixed excitation could be treated by itself contexts, the mixed excitation could be treated by itself as a different class. The non-voiced region includes silence and unvoiced speech [1]. The voiced and unvoiced section can be classified using these features. Speech is time-varying and the model parameters are also time-varying so short-time analysis to estimate is needed. Furthermore, from speech samples to model parameters, alternative short-time representations are often required.

II. SPEECH ANALYSIS BY WINDOWING

The properties of speech signal change relatively slowly with rates of change on the order of 10 - 30 times per sec, corresponding to the rate of speech 5 - 15 phones or sub phones per second. A speech

signal is partitioned into short segments, each of which is assumed to be similar to a frame from a sustained sound. Such a segment is called a frame. The frames are used to detect the sounds, are then integrated to be the speech. Window function $w[n]$ is used to extract a frame from the speech waveform. The commonly used windows are the rectangular and Hamming. The equation of the following windows are defined as

$$w_R[n] = 1, \quad 0 \le n \le L-1$$
$$= 0, \quad otherwise \qquad (1)$$

$$w_H[n] = 0.54 - 0.46\cos\left(\frac{2\Pi n}{L-1}\right), \quad 0 \le n \le L-1$$
$$= 0, \quad otherwise \qquad (2)$$

Where, L is the length of a frame. The resolution offered by the rectangular window function is better in comparison with Hamming window as the width of main lobe of rectangular window is smaller [2]. Relatively there is more spectral leakage in case of rectangular window as the peak-to-side lobe ratio is low which is not desirable in comparison with Hamming window. Thus from the resolution point of view, rectangular window is preferable and from spectral leakage point of view Hamming window are preferable. The effect of spectral leakage is very severe and it affects the speech signal to be analysed, hence Hamming window is employed.

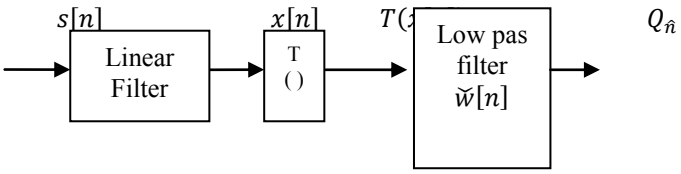

Figure 1.Short time processing

All the short-time processing can be represented mathematically as in Eq.3

$$Q_{\hat{n}} = \sum_m T(x[m])\tilde{w}[\hat{n} - m] \qquad (3)$$

T (·) is meant to extract certain feature of the speech signal. The feature(s) is then summed over a window $\tilde{w}[\hat{n} - m]$anchored at $\hat{n}$.

III. SPEECH ANALYSIS IN TIME DOMAIN

3.1. Short time energy and short time magnitude

The amplitude of unvoiced segments is very lower than the amplitude of voiced segments[2]. The short time energy of the speech signal provides convenient representation that reflects these amplitude variations. The short time energy is defined in Eq.6 as

$$E_{\hat{n}} = \sum_m (x[m]w[\hat{n} - m])^2 \qquad (4)$$

One disadvantage of short time energy is that it is very sensitive to large signal, thereby emphasizing large sample to sample variations the short time magnitude is defined in Eq.7 as

$$M_{\hat{n}} = \sum_m |x[m]w[\hat{n} - m]| \qquad (5)$$

Short time magnitude is similar to short time energy where the weighted sum of absolute values

of the signal is computed instead of sum of the squares [3].Short time energy and short time magnitude is useful in detecting voiced segments of speech. It is also useful to detect silence segments. The energy and magnitude is high in voiced section and less in unvoiced section and very less almost zero in silence section of the speech signal [2]. Short time magnitude computation is easier than short time energy.

3.2. Short time zero crossing rate

A zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. The ZCR in case of stationary signal is defined in Eq.8

$$Z_n = \sum_{m=-\infty}^{\infty} |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]| w(n-m) \quad (6)$$

Where sgn (s(n))=1 if s(n)≥0
=-1 if s(n)<0

This relation can be modified for non stationary signals like speech and termed as short time ZCR. It is defined in Eq.9

$$z(n) = 1/2N \sum_{m=0}^{N-1} s(m).w(n-m) \quad (7)$$

The factor 2 is because there will be two zero crossings per cycle of one signal. Short time ZCR is used for detecting the voiced and unvoiced section[4]. It can be also used for end point detection or silence removal. A voiced section is low in zero crossing rates and unvoiced is medium in zero crossing rates and highest in silence section [2].

3.3. Short time autocorrelation

The deterministic autocorrelation function of a discrete-time signal x[n] is defined in Eq.10

$$\emptyset[k] = \sum_{-\infty}^{\infty} x[m]x[m+k] \quad (8)$$

At analysis time $\hat{n}$ the short-time autocorrelation is defined as the autocorrelation function of the windowed segment as in Eq.11

$$R_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n}-m)(x[m+k]w[\hat{n}-k-m]) \quad (9)$$

It is used for voiced and unvoiced section decision .If STACR close to being impulse then it is unvoiced and if STACR periodic with tapered amplitude then it is voiced. It is also used for pitch detection.

3.4. Short time average magnitude difference function

An alternative to the autocorrelation is the average magnitude difference function (AMDF). Rather than multiplying speech x (m) by x (m − k), the magnitude of their difference is used as in Eq.12

$$AMDF(k) = \sum_{-\infty}^{\infty} |x(m) - x(m-k)| \quad (10)$$

Subtraction is a simpler computer operation than multiplication hence AMDF is much faster.

IV.SPEECH ANALYSIS IN FREQUENCY DOMAIN

4.1. Short time Fourier transforms

Time varying spectral information is taken into account hence, the short time processing approach is employed. In short term processing, speech is processed is blocks of 10-30 ms with a shift of 10 ms. To accommodate the time varying nature of this spectrum, the DTFT equation is defined as Eq.13

$$X(w,n) = \sum_{-\infty}^{\infty} x(m)w(n-m)e^{-jwn} \qquad (11)$$

Where, *W(n)* is the window function for short term processing. The spectral amplitude and phase are function of both frequency and time where as it was only function of frequency in the earlier case of DTFT. *x(m).w(n-m)* represents the window segment around the time instant 'n'**.** Hence *X(w,n)* at *'n'* represents the spectrum of the speech segment present around it. When *'n'* is shifted, then correspondent *X(w,n)* also changes[6]. Thus showing the time varying spectra of speech.[3] Since such a time-spectral is computed using short term processes, *X(w,n)* is termed as Short Term Fourier Transform (STFT).

4.2. Spectrogram

For any specific window type, its duration varies inversely with spectral bandwidth, i.e., the usual compromise between time and frequency resolution [2]. Wideband spectrograms display detailed time. Narrow band spectrograms typically use a 20 ms with a corresponding 45 Hz bandwidth, ,thus they display individual harmonics but the time-frequency representation undergoes significant temporal smoothing [6]. In Narrow band spectrogram since L is increased the bandwidth is decreased. It provides good frequency resolution and bad temporal resolution. It is used for pitch estimation. In wideband spectrogram since L is decreased the bandwidth is increased. It provides good temporal resolution and bad frequency resolution. It is used for viewing vocal tract parameters which can change slowly and hence do not need fine frequency resolution.

VI. RESULT AND CONCLUSION

The time domain and frequency domain algorithms were implemented using MATLAB R2009a and the interfaced with graphic user interface (GUI). The interfaced GUI model is as shown below

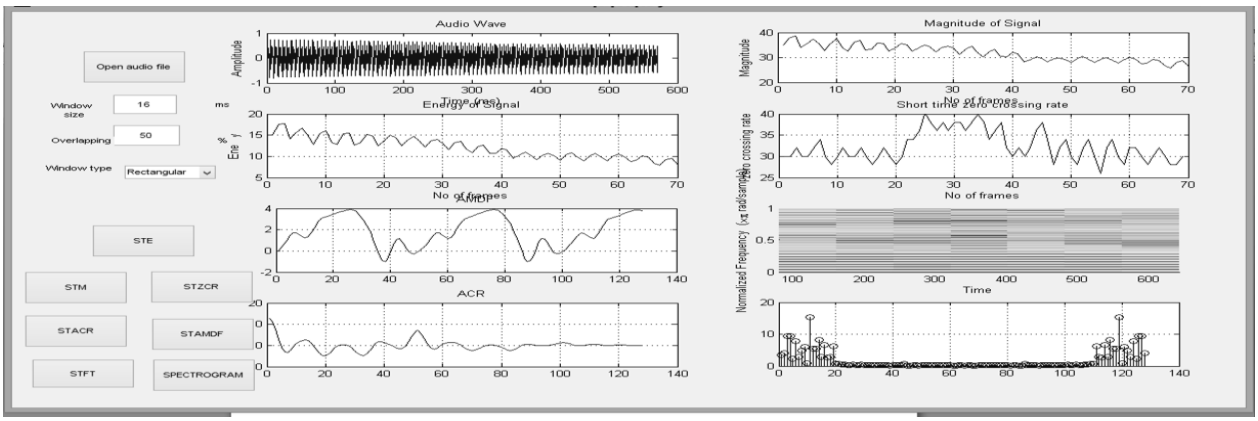

Figure 2.GUI model showing results of algorithms of voiced speech (/a/)

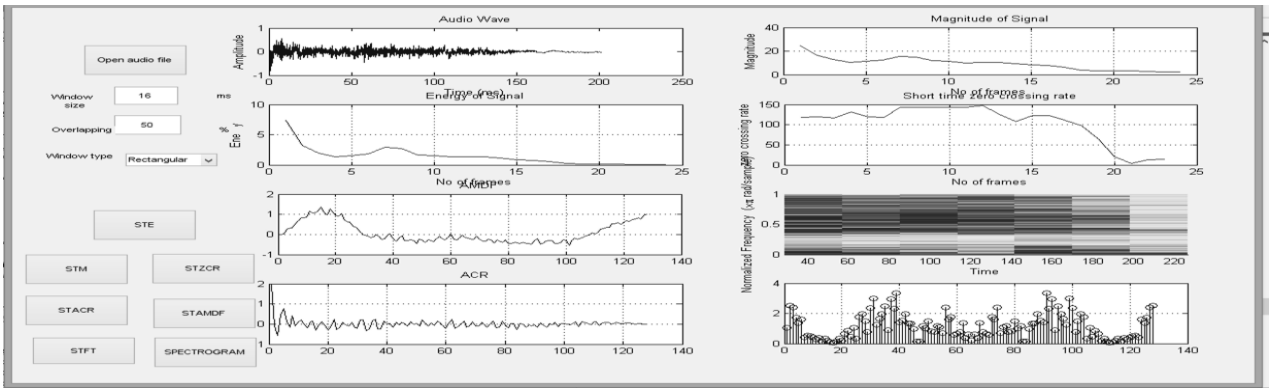

Figure 3.GUI model showing results of algorithms of unvoiced speech (/sh/)

Short time energy and magnitude is useful in detecting voiced segments of speech. Short time zero crossing rate with energy, can be used in the classification of voiced/unvoiced segments of speech signal. A voiced segment is low in zero crossings rate, medium in unvoiced section and high in silence section [5]. It is also used to detect the silence region of the speech. For voiced segments, the autocorrelation function shows periodicity. Pitch can be estimated using autocorrelation function. Short time average magnitude difference function can be used for voiced unvoiced decision. It is more efficient as it uses difference instead of multiplication. As the window length increases, short-time energy and magnitude becomes smoother. The results observed from the GUI model are as follows.

Table 1.Comparison of algorithms in voiced and unvoiced speech

	ENERGY	*MAGNITUDE*	*ZERO CROSSING RATE*	*AUTOCORRELATION*	*AVERAGE MAGNITUE DIFFERENCE EQUATION*	*STFT*
VOICED	HIGH	MEDIUM	LOW	PERIODIC	PERIODIC	PERIODIC
UNVOICED	LOW	HIGH	MEDIUM	APERIODIC	APERIODIC	APERIODIC

REFERENCES

[1] D. O'Shaughnessy, *Speech Communications: Human &Machine,* 2nd ed. Universities Press India Limited, India,2001.
[2] L. R. Rabiner and R. W. Schafer, "*Digital Speech Processing of Speech Signals*" 3rd ed. Pearson Education, 2009, pp. 132-174.
[3] Ghulam Muhammad "Extended Average Magnitude Difference Function Based Pitch Detection" *in Proceedings of The International Arab Journal of Information Technology* Vol.8, pp 197-203, April 2011.
[4] Bhargab Medhi and P.H.Talkudhar "Assamese Vowel Phoneme Recognition Using Zero Crossing Rate and Short-time Energy" *in Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering,* Vol 4, Issue 4 April 2014.
[5] Ykhlef Faycel and Messaoud Bensebti "Compartive Performance for Voiced/Unvoiced Classification" *in Proceedings of International Arab Journal of Information Technology*, Vol 11, Issue No 3,May 2014.
[6] Xinglei Zhu and Gerlad T.Beauregard "Real Time Signal Estimation from Modified Short Time Fourier Transform Magnitude Spectra" *in Proceedings of IEEE Transaction on audio, speech and language processings,* Vol .15,Issue No.5,July 2007 .

