

Discriminating Voiced and Unvoiced Segments of Speech Signal using GUI

Solly Joy Savitha Upadhya

Abstract- In this paper, the concepts of speech processing algorithms for speech signal analysis is presented using the GUI model of matlab. 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. SHORT TIME PROCESSING

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 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

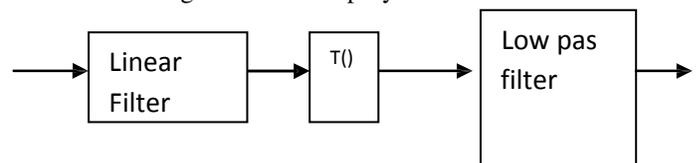
$$w_R[n] = 1, \quad 0 \leq n \leq L-1$$

$$= 0, \quad \text{otherwise} \quad (1)$$

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

$$= 0, \quad \text{otherwise} \quad (2)$$

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.



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

(3)

$T(\cdot)$ is meant to extract certain feature of the speech signal. The feature(s) is then summed over a window anchored at

III. TECHNIQUES IN TIME DOMAIN

A. 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.4 as

$$= \quad (4)$$

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.5 as

(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.

B. 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.6

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

$$\text{Where } \text{sgn}(s(n)) = 1 \text{ if } s(n) \geq 0 \\ = -1 \text{ if } s(n) < 0$$

This relation can be modified for non stationary signals like speech and termed as short time ZCR. It is defined in Eq.7

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

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].

C. Short Time Autocorrelation

The deterministic autocorrelation function of a discrete-time signal $x[n]$ is defined in Eq.8

(8)

At analysis time the short-time autocorrelation is defined as the autocorrelation function of the windowed segment as in Eq.9

(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.

D. Short Time Average Magnitude Difference Magnitude

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.10

(10)

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

IV. TECHNIQUES IN FREQUENCY DOMAIN

A. 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.

B. 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 in 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.11

$$(11)$$

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).

V. SPEECH SEGMENTS EVALUATION

The algorithms are implemented using MATLAB R2009a. Audacity is used for recording sounds. Audacity supports audio files with .wav extensions. These recorded sounds were read in MATLAB, using wavread command. Then various algorithms were applied on these signals to discriminate voiced and unvoiced segments of speech. Then the algorithms implemented are displayed using Graphic user interface .The GUI is interfaced with MATLAB using GUIDE. It provides analysis of speech signal with varying window size and shape and overlapping also.

VI. RESULTS

The GUI model displays the analysis techniques for classification of voiced and unvoiced segments of speech signals. Fig.2 shows the graphical results of the algorithms in time and frequency domain. Energy is high in voiced section. Magnitude is high in voiced region. Magnitude of speech signals becomes smoother by increasing window size. This helps in differentiating between voiced and unvoiced section of speech signal. As the window size is increased the energy becomes smoother while at smaller window ,energy fluctuates a lot. Zero crossing rate is low. Autocorrelation and AMDF is periodic in voiced section. This periodicity is used for estimating pitch of the speech signal. ZCR can be used in end point detection. STFT is periodic in voiced section which is used in feature extraction techniques.

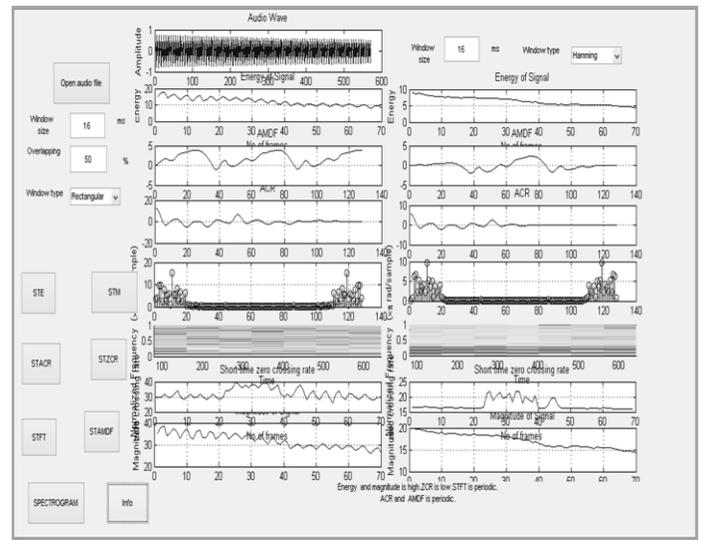


Fig .2 GUI model representing voiced sample(/a/) using Hamming and Rectangular window

Fig.2 shows the graphical results of the algorithms in time and frequency domain. Energy is low in unvoiced section[8]. Magnitude is also low in unvoiced section. This helps in differentiating between voiced and unvoiced section of speech signal. Zero crossing rate is high. Autocorrelation and AMDF is non periodic in unvoiced section. . STFT is nonperiodic in unvoiced section .All the information are displayed in the GUI model.

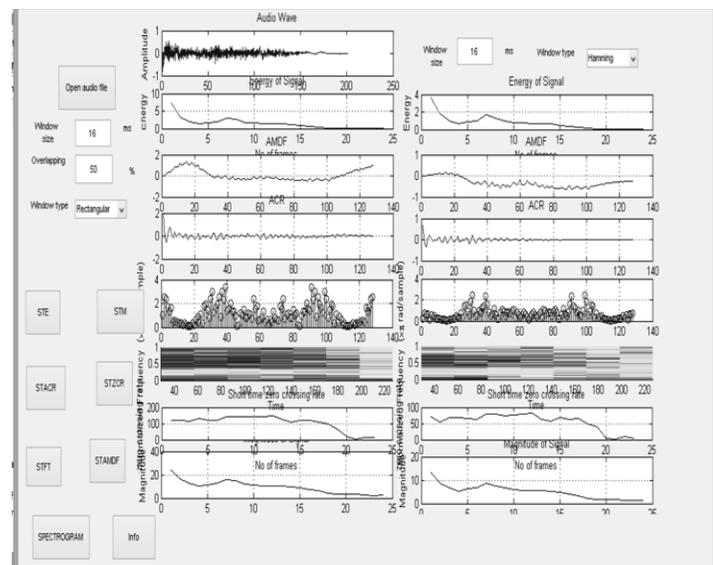


Fig .3 GUI model representing unvoiced sample(/sh/) using Hamming and Rectangular window

VII. CONCLUSION

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 I. COMPARISON OF ALGORITHMS IN VOICED AND UNVOICED SPEECH

Technique	Voiced	Unvoiced
Energy	High	Low
Magnitude	High	Low
Zero crossing rate	Low	High
Average magnitude difference function	Periodic	Impulse
Autocorrelation	Periodic	Impulse
STFT	Periodic	Non periodic

From the tables it is observed these techniques can be used to differentiate voiced and unvoiced segment of speech signal. Short time energy and Short time zero crossing rate together is used for end point detection which is useful in speech recognition.

REFERENCES

- [1] D. O'Shaughnessy, *Speech Communications: Human & Machine*, 2nd ed. Wiley-IEEE Press, 1999, pp. 367-435.
- [2] L. R. Rabiner and R. W. Schafer, "Digital Speech Processing for Man-Machine Communication by Voice" in *Digital processing of Speech Signals*, 3rd ed. Pearson Education, 2009, pp. 505-516
- [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 "Comparative 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 Transactions on audio, speech and language processings, Vol .15, Issue No.5, July 2007
 [7] 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 and Engineering*.