Speech Analysis in Time and Frequency Domain

Solly Joy
EXTC Department
FCRIT
Vashi,Navi Mumbai,India

Savitha Upadhya
EXTC Department
FCRIT
Vashi,NaviMumbai,India

***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 energy, Short time magnitude, Short time zero crossing rate, Spectrogram.**

I INTRODUCTION

Most speech processing applications utilize certain properties or features of speech signals in accomplishing their tasks. The extraction of these properties or features and how 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 (vibration of the vocal folds). 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 (mixed excitation, such as voiced fricatives), 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, fromspeech samples to model parameters, alternative short-time representations are often required. Short time analysis is also known as windowing.[2]. The speech signal is segmented and multiplied with the window function. Short time analysis provides better results than the complete speech signal analysed .The short time domain analysis are energy, magnitude, autocorrelation and average magnitude difference function. The frequency domain analysis are Short time Fourier transform, Wide band spectrum, Narrow band spectrum.

II SHORT TIME ANALYSIS

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, which are integrated to be the speech. Window function $w[n]$ is used to extract a frame from the speech waveform. There are different types of frames such as short frames (5 − 20 ms), medium frames (20 − 100 ms), long frames (100 − 500 ms).The commonly used windows are the rectangular and Hamming windows as. The equation of these windows

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

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

Where L is the length of a frame.

The width of main lobe for rectangular window is small compared to the Hamming window. As a result the resolution offered by the rectangular window function is better. The peak-to-side lobe ratio of rectangular window is significantly poor compared to the Hamming window[2]. This results in relatively more spectral leakage in case of rectangular window which is not desirable. 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 severe so it affects speech signal analysis, hence Hamming window is employed.

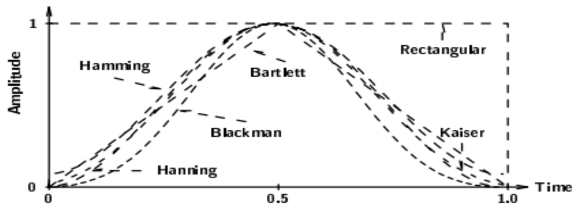

Fig. 1. Common windows

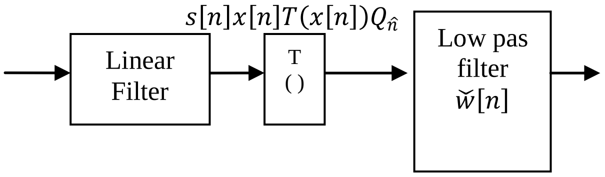

Fig. 2.General representation of short time analysis

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

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

T (·) is meant to extract certain feature(s) of the speech signal. The feature(s) is then summed over a window$\tilde{w}[\hat{n} - m]$anchored at$\hat{n}$. The result is a short-time feature near$\hat{n}$. For example, for a rectangular window with length L,

$$Q_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} T(x[m]) (4)$$

Let $\hat{n}$ be shifted R samples a time,

$\hat{n}$=kRk = 0, 1 (5)

where R is the time between frames.

The choice of R is dependent on the frame length L.

III TIME DOMAIN ANALYSIS

A. Short Time Energy and Magnitude

The amplitude of unvoiced segments is generally much 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

$$E_{\hat{n}} = \sum_m (x[m]w[\hat{n} - m])^2 \quad (6)$$

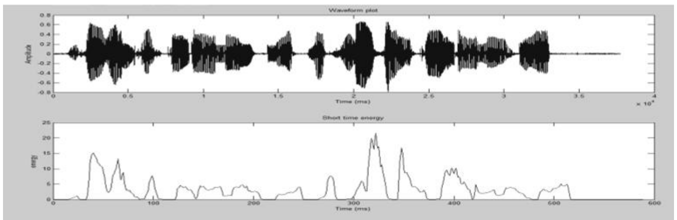

Fig.3.Short time energy for 16ms window size and rectangular window is used.

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

$$M_{\hat{n}} = \sum_m |x[m]w[\hat{n} - m]| (7)$$

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].

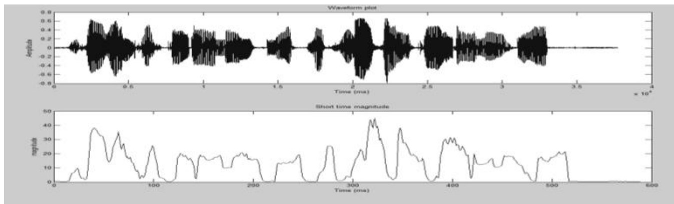

Fig 4.Short time magnitude for window size 16ms using rectangular window.

Short time energy is useful in detecting voiced segments of speech. It is also useful to detect silence segments. The energy 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 is useful in detecting voiced and unvoiced section and computation is easier than energy. Magnitude is high in voiced section and low in unvoiced [4].

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.8

$$Z_n = \sum_{m=-\infty}^{\infty} |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]| w(n-m) \quad (8)$$

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

$$z(n) = 1/2N \sum_{m=0}^{N-1} s(m).w(n-m) \quad (9)$$

The factor 2 is because there will be two zero crossings per cycle of one signal.

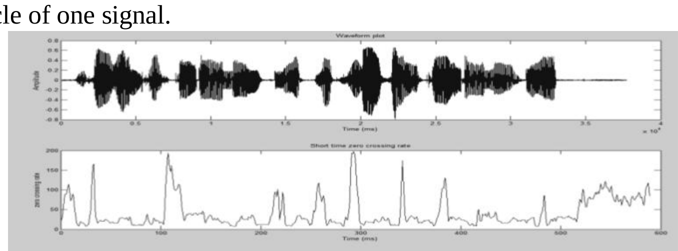

Fig 5.Short time zero crossing rates for window size 16ms using rectangular window.

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.10

$$\emptyset[k] = \sum_{-\infty}^{\infty} x[m]x[m+k] \quad (10)$$

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

$$R_{\hat{n}}[k] = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n}-m)(x[m+k]w[\hat{n}-k-m]) \quad (11)$$

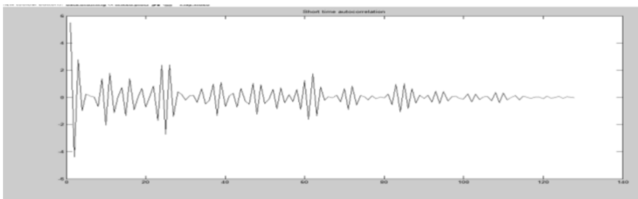

Fig 6.Short time autocorrelation for window size 16ms using rectangular window.

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 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

$$AMDF(k) = \sum_{-\infty}^{\infty} |x(m) - x(m-k)| \quad (12)$$

Assuming that subtraction is a simpler computer operation than multiplication, the AMDF is faster. The AMDF has mimima for values of k near multiples of pitch period.

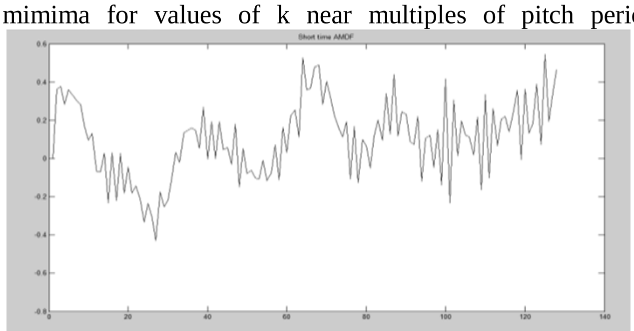

Fig 7.Short time average magnitude difference function for window size 16ms using rectangular window.

Short time AMDF gives better result as it is easy for computation.It is used for voiced unvoiced section decision.It can be used for pitch detection [5].

IV FREQUENCY DOMAIN ANALYSIS

Short Time Fourier Transform

To take care of time varying spectral information, 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. For instance, using a block size of 20 ms, the DTFT is computed using DFT for that block. Their process is repeated for all the blocks of speech signal and all the spectra computed are stacked together as a function of time and frequency to observe the time varying spectra. To accommodate the time varying nature of this spectrum, the DTFT equation is defined as Eq.13

$$X(w,n) = \sum_{-\infty}^{\infty} x(m)w(n-m)e^{-jwn} \quad (13)$$

where *W(n)* is the window function for short term processing. Now 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 giving visualization of 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).

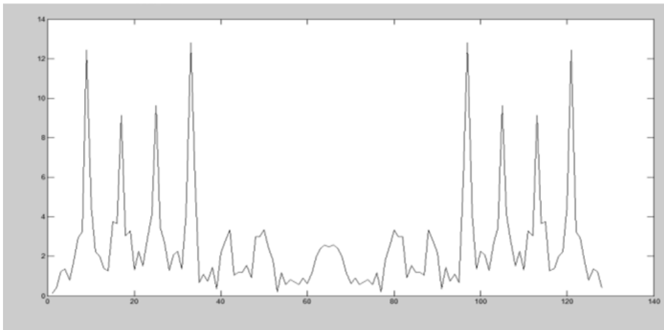
Fig 8.Log magnitude spectra of /a/

A. Wide band and Narrow band 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. (Amplitude variations corresponding to vocal cord closures), and typically use a window about 3 ms long. This means a bandwidth of approximately 300 Hz, which smoothes away harmonic structure (except for very high-pitched voices). 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].

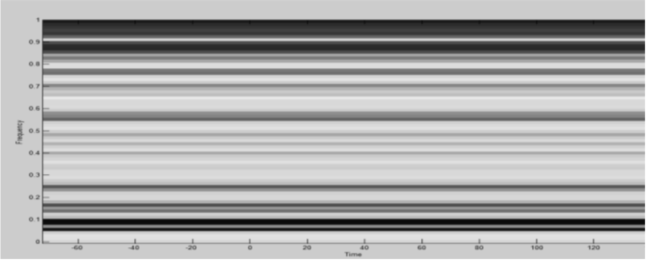

Fig 9.Wide band spectrogram where window size is 3 ms andrectangular window is used.

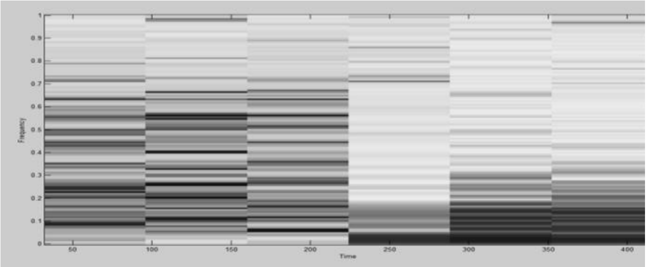

Fig 10 .Narrow band spectrogram window size is 20 ms and using rectangular window.

In Narrow band spectrogram sinceL isincreased the bandwidth is decreased. It provides good frequency resolution and bad temporal resolution. It is used for pitch estimation. In wideband spectrogram sinceL 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.

V Conclusion

Short time energy is useful in detecting voiced segments of speech. It is also useful to detect silence segments. It is observed that the energy is high in voiced section and less in unvoiced section .Short time magnitude is also useful in detecting voiced and unvoiced segments of speech [2]. 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.

TABLE 1.COMPARSION OF MAGNITUDE AND ZERO CROSSING RATE FOR VARIOUS SPEECH SEGMENTS.

	VOICED	*UNVOICED*	*SILENCE*
M_N	HIGH	MEDIUM	LOW
Z_N	LOW	HIGH	MEDIUM

VI FUTURE SCOPE

The software is developed using MATLAB R2009a.GUI interface of the same could be implemented which will display the parameters of speech signal to be analysed. It will display the time domain and frequency domain parameters of a speech signal which can then be analysed and studied.

References

[1] O'Shaughnessy,*Speech Communications: Human & Machine*,2nd ed. Universities Press India Limited, India,2001.
[2] L. R. Rabiner and R. W. Schafer,"Digital Speech Processingof Speech Signals"3rded. 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 TechnologyVol.8,pp 197-203, April 2011.
[4] BhargabMedhi and P.H.Talkudhar"Assamese Vowel Phoneme Recognition Using ZeroCrossing Rate and Short- time Energy" in Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Vol 4,Issue4April 2014.
[5] YkhlefFaycel and MessaoudBensebti "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 GerladT.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