International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 4, April 2015
Available at http://internationaljournalofresearch.org

Computer Software Package for Analysis of Speech Signal

Solly Joy
EXTC Department. FCRITVashi, Navi Mumbai, India
sollyjoy42@gmail.com
Savitha Upadhya
EXTC Department. FCRITVashi, Navi Mumbai, India
savivasshi@gmail.com

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 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. This is also known as frame by frame analysis or windowing.

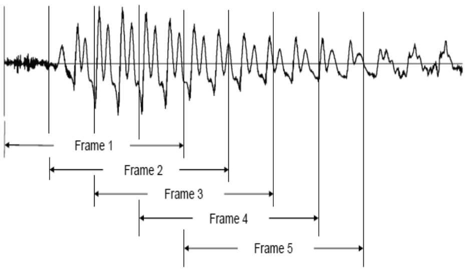


Fig.1.Frame by Frame analysis

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 Rectangular and Hamming 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.

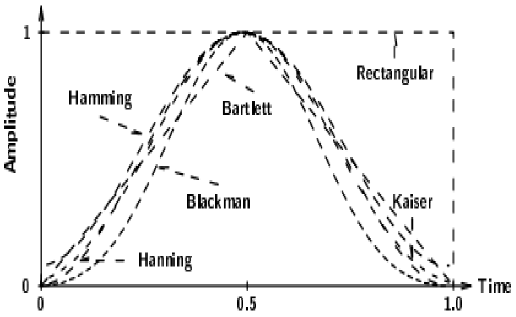


Fig.2.Commonly used windows

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

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

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

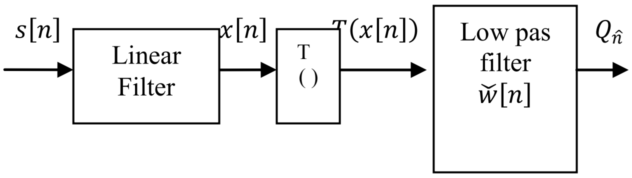


Fig.3.Short time processing.

III.TECHINQUES 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

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

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

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\right| w(n-m)$$
(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.7

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

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

At analysis time $\hat{n}$ the short-time autocorrelation is defined as the autocorrelation function of the windowed segment as in Eq.9

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

$$AMDF(k) = \sum_{-\infty}^{\infty} |x(m) - x(m-k)|$$
(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 variation .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[5].

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 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.11

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

C. *Cepstrum*

According to the source filter theory of speech production, voiced sounds are produced by exciting the time varying system characteristics with periodic impulse sequence and unvoiced sounds are produced by exciting the time varying system with a random noise sequence. The resulting speech can be considered as the convolution of respective excitation sequence and vocal tract filter characteristics. If e(n) is the excitation sequence and h(n) is the vocal tract filter sequence, then the speech sequence s(n) can be expressed as follows:

$$s(n) = h(n) * e(n)$$
(12)

This can be represented in frequency domain as,

$$s(\omega) = h(\omega).e(\omega)$$
(13)

The above equation indicates that the multiplication of excitation and system components in the frequency domain for the convolved sequence of the same in the time domain. The speech sequence has to be deconvolved into the excitation and vocal tract components in the time domain. For this, multiplication of the two components in the frequency domain has to be converted to a linear combination of the two components. For this purpose cepstral analysis is used for transforming the multiplied source and system components in the frequency domain to linear combination of the two components in the cepstral domain. Ceostrally smoothened is the smoothened version of cepstrum and is obtained by taking the FFT of the cepstrum c(n).

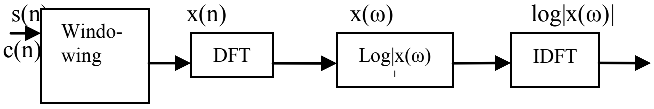


Fig.4.Block diagram representing computation of cepstrum.

D. Pitch and Formants

Pitch F0 also known as fundamental frequency is the relative highness or lowness of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the vocal cords[7]. Pitch is the main acoustic correlate of tone and intonation. Pitch is detected with help of methods like autocorrelation, cepstrum, average magnitude difference method.The spectral peaks of the sound spectrum are called formants. Formants are the distinguishing or meaningful frequency components of human articulation[8]. The formant with the lowest frequency is called F_1, the second F_2, and the third F_3.

V. METHOD

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 AND DISCUSSIONS

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 in voiced section. 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.

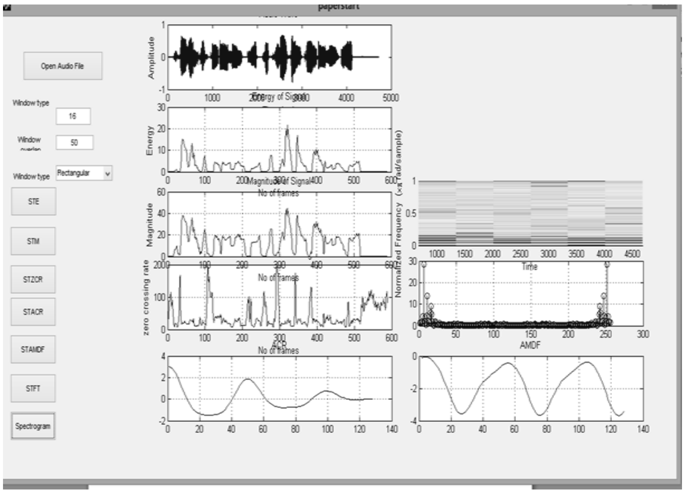


Fig.5GUI model representing speech sample(/machali/) using rectangular window

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 in unvoiced section. Autocorrelation and AMDF is non periodic in unvoiced section. . STFT is nonperiodic in un voiced section .All the information are displayed in the GUI model.

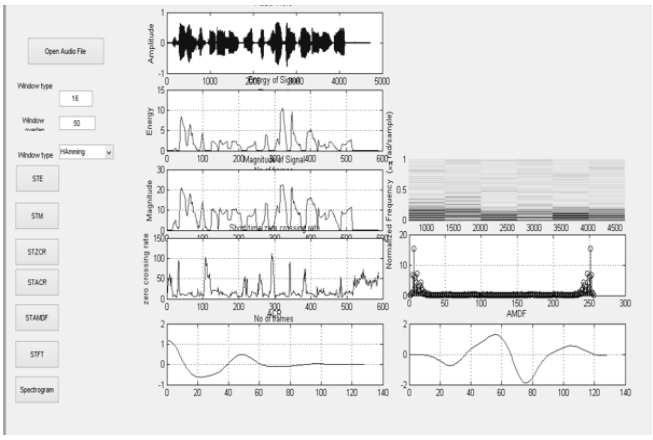
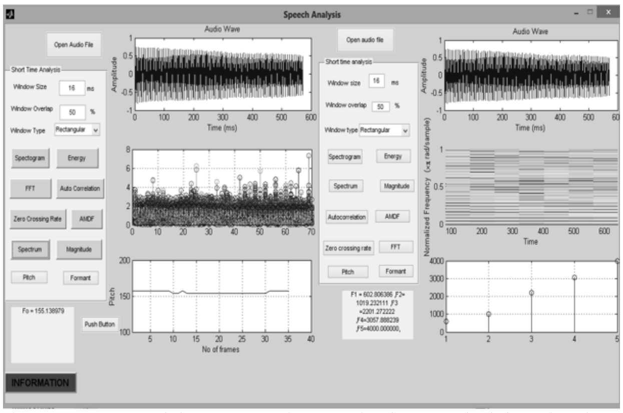


Fig.6.GUI model representing speech sample(/machali/) using Hamming window



Window Size 16 ms

Window Overlap 50 %

Window Type Rectangular

Fo = 155.138979

F1 = 602.806386 ,F2= 1019.232111 ,F3 =2201.272222 ,F4=3057.888239 ,F5=4000.000000,

Fig.7 GUI model representing voiced sample(/a/) showing pitch,formants and cepstum.

Cepstrum is symmetrical in voiced section. Pitch is constant over the entire speech signal.For men and women the size difference of the vocal folds, reflecting male-female differences in larynx size, will influence pitch range so that adult male voices are usually lower-pitched with larger folds than female voices[7]. In speech, pitch helps to identify the gender of the speaker (pitch tends to be higher for females than for males), ages (adults or children) and gives additional meaning to words and may help to identify the emotional state of the speaker (e.g., joy produces high pitch and a wide pitch range, while sadness produce normal to low pitch and a narrow pitch range[9]. Here the vowel taken is /a/ and is recorded by male hence the pitch obtained is 152 Hz

Most often the two first formants, F_1 and F_2, are enough to disambiguate the vowel. Thus the first formant F_1 has a higher frequency for an open vowel (such as [a]) and a lower frequency for a close vowel (such as[i] or [u]); and the second formant F_2 has a higher frequency for a front vowel (such as [i]) and a lower frequency for a back vowel (such as [u]).Vowels will almost always have four or more distinguishable formants; sometimes there are more than six[8].The formants value obtained are 271Hz,991Hz and 1922Hz. The data is also displayed in the GUI model for better analysis of speech signal in time and frequency domain.

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. In unvoiced segment the variations in the lower quefrency region (near 0 axis) is due to vocal tract characteristics and the fast varying nature of the cepstrum towards the upper quefrency region represents the excitation characteristics of the short term speech segment and symmetric in voiced section in cepstral analysis. Cepstral analysis is used for determining pitch as well as formant. The first two formants are most important in determining vowel quality. The fundamental frequency of speech can vary from 40 Hz for low-pitched male voices to 600 Hz for children or high-pitched female voices[8].The results observed from the GUI model are as follows

TABLE I. COMPARISON OF ALGORITHMS IN VOICED AND UNVOICED SPEECH

Techinque	***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

VIII.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 "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 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"*IEEEInternationalJoint Conferences on Computer,Information,and Systems and Engineering.*

[8]Sassan Ahmadi and Andreas S.Spanias"Cepstrum Based Pitch Detection Using a New Statistical V/UV Clasification Algorithim" *in Proceedings of IEEE Transactions on speech and audio processings,*Vol .7,Issue No.3,May 1999 .

[9]Bageshree V .Sathe Pathak and Ashish R.Panat"Extraction of Pitch and Formants and its Analysis to Identify 3 Different Emotional States of a Person" *in Proceedings of IJCSI International Journal of Computer Science* Vol .9,Issue No.1,July 2012 .