Automatic Identification of Silence, Voiced and Unvoiced Chunks in Speech

A Thesis

Submitted in partial fulfillment of the requirements for the award of the degree of

Master of Technology

Submitted by

Poonam Sharma
(Roll No. 601003019)

Under the supervision of

Dr. R. K. Sharma
Professor
School of Mathematics and Computer Applications
Thapar University
Patiala

School of Mathematics and Computer Applications
Thapar University
Patiala – 147004 (Punjab), INDIA

June 2012

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, **"Automatic Identification of Silence, Voiced and Unvoiced chunks in Speech"**, in partial fulfillment of the requirements for the award of degree of Master of Technology in **Computer Science and Applications** submitted in School of Mathematics and Computer Applications of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Dr. R.K. Sharma and refers other researcher's work which are duly listed in the reference section.

The matter presented in this thesis has not been submitted for award of any other degree of this or any other University.

Poonam Sharma.

(Poonam Sharma)

Roll No.: 601003019

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

R K Sharma 12.6.12

(Dr. R. K. Sharma)
Professor, SMCA
Thapar University, Patiala

Countersigned:

(Dr. S.S. Bhatia)
Head,
School of Mathematics & Computer Applications
Thapar University,
Patiala

(Dr. S. K. Mohapatra)
Dean of Academic Affairs
Thapar University,
Patiala

ABSTRACT

Computers are greatly influencing the lives of human beings and their usage is increasing at a tremendous rate. The ease with which we can exchange information between user and computer is of immense importance today. But the input devices like mouse and keyboard have their limitations when used as an interface to exchange the information. Speech which is natural and quick way of exchanging the information between humans, if used to communicate with computers can overcome all these limitations. Speech recognition is in research for many years and has attracted many researchers across the world. Detection of word boundary, silence detection, voiced unvoiced detection, noise removal, effects of voice quality are the prominent problems for achieving high degree of accuracy in speech recognition. The main goal of this thesis is to design an algorithm for automating the detection of silence, voiced and unvoiced chunks in speech signal which is very important for increasing accuracy of any recognition system. This thesis is divided into five chapters. A brief outline of each chapter is given in the following paragraphs.

Chapter 1 includes two sections namely, speech recognition and its issues and literature survey. Issues in speech recognition include: Silence, unvoiced and voiced detection, noise and voice quality. In literature review a detailed literature survey on the algorithms and methods used until now for word boundary detection and silence, unvoiced and voiced classification is done chronologically.

Chapter 2 contains the work carried out for the three important phases namely data collection, preprocessing and feature extraction for the automation of the classification. In data collection phase sounds were recorded of 3 males and one female member. 15 words were spoken 3 times by each member in Hindi. After that in preprocessing windowing of the speech signal was done using rectangular and hamming window and than three different features namely, zero crossing rate, short time energy and fundamental frequency were calculated which were used for the automation of the algorithm.

Chapter 3 focuses on the main work done for automation. Its first section discuss the results that are derived from the calculation of feature vectors and are helpful in the identification of the silence, voiced and unvoiced chunks in the input speech signal. The next section describes

how these results are used to develop the algorithm and provides the information regarding the steps that are followed for automation and the corresponding flow chart.

Chapter 4 discusses the outputs of the algorithm showing graphs for different words showing the situations when the algorithm was almost completely identifying the signal correctly and when it was not in some situations. The overall accuracy of the algorithm is found out to be 96.61 %.

Chapter 5 presents the conclusion of the work done and also the advancements that can be made to the work to increase the accuracy.

ACKNOWLEDGEMENTS

My sincere thanks to all the people around me who helped me in completing this thesis work. First, I wish to thank **Dr. R. K. Sharma** (Professor) of School of Mathematics and Computer Applications, Thapar University, Patiala for giving me an opportunity to work under his guidance. His continued support, guidance and vision helped me to complete this thesis. It has been a pleasure working under his guidance.

I truly appreciate cooperation and support received from my friends Rahul Aggrawal, Ajay Sharma and Mayank Gupta, during this work.

I also express my sincere gratitude to all the faculty members at **THAPAR UNIVERSITY** for equipping me with the best of knowledge and providing me top class facilities and infrastructure.

Date: 11th June 2012

Poonam Sharma

Place: Thapar University, Patiala. (Poonam Sharma)

LIST OF FIGURES AND GRAPHS

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Expanded Form
MFCC	Mel Frequency Cepstral Coefficients
LPC	Linear Predictive Coding
LDM	Linear Data Modulation
MiMSB	Minimum Mel Scale Frequency Band
ETF	Enhanced Time Frequency
ZCR	Zero Crossing Rate
STE	Short Time Energy
F0	Fundamental Frequency

CONTENTS

CHAPTER 1

INTRODUCTION

Since the computers have been evolved we are dealing with various research activities in the area of human computer interface. The input devices such as keyboard and mouse are although very popular mediums to interact with the computer but has some limitations as keyboard requires a certain amount of skill for effective and fast usage and mouse on the other hand requires a good hand and eye coordination. The physically challenged people find computers difficult to use.

Speech which is a natural and very easy way of exchanging the information if used as a medium to interact with the computer and can solve all these problems. Speech recognition technology has made it possible for computers to follow human voice commands and understand human languages. The main goal of speech recognition area is to develop techniques and systems for speech as input to machine.

From past 60 years many researches and advancements have taken place in this area, many systems have been developed but still after years of research and development the accuracy of automatic speech recognition remains one of the important research challenges. From Speech representation, Speech classification to recognition and performance evaluation there are still many issues that are to be handled.

1.1 SPEECH RECOGNITION AND ITS ISSUES

Speech Recognition also known as automatic speech recognition is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program or we can say it is the ability of the computer to accept speech in audio format and then generate its content in text format. Typical model for Speech Recognition is given in Figure 1.

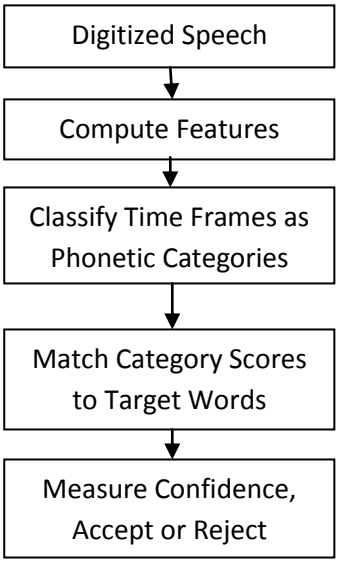

Figure 1: Typical speech recognition model.

Speech recognition in computer domain includes various steps with issues attached with them. The general model begins with a user creating a speech signal which is amplitude versus time waveform. This digitized speech signal is used to extract various spectral and temporal features like zero crossing rate, short time energy, fundamental frequency, mfcc *etc.* Some of these features are used for word boundary detection, silence detection *etc.* which are done during the preprocessing of the speech signal and many along with these are used for recognition in subsequent phases by making a feature vector. These feature vectors are compared against stored and trained knowledge model to categorize phonemes which are further combined to form the target words. These words depending upon their probabilistic confidence either are accepted or rejected.

Although there are many issues which makes the speech recognition a challenging task but the main issues among them are noise removal, word boundary detection, voice quality detection *etc*.

1.1.1 Detecting voiced, unvoiced and silence region

Speech signal is a slowly time varying signal when examined over a small interval of time but when speech is being taken for large time its features are not stationary. There are several ways of classifying or labeling the speech signal but the simplest and the most straightforward way is via the state of the speech production. Generally a three state representation is used consisting of following stages.

Silence: In this state of speech, no sound is being produced so the energy and the amplitude of the signal is very low. This is important to identify silence region. Once identified than that part of the speech signal can be ignored for further recognition process.

Unvoiced: In this stage of speech, vocal cords do not vibrate so the resulting speech is random in nature like the sounds of whisper or aspiration. Also for fricatives (*e.g.*, /f/ as in fish or /s/, as in mess or /sh/, as in shalgam), unvoiced excitation (noise) is used. In these cases, usually no fundamental frequency can be detected. On the other hand, the zero crossing rate of the signal is very high.

Voiced: In this stage of speech, vocal cords are tensed and vibrate periodically. Voiced excitation for the speech sound will result in a pulse train called as fundamental frequency. Voiced excitation is used when articulating vowels and some of the consonants.

Segmentation of the waveform into well defined regions of silence, unvoiced and voiced regions is not exact as it is difficult to distinguish a week unvoiced sound (like /f/) from silence or a week voiced sound (like /v/) from unvoiced. This segmentation is being illustrated in Figure 2.

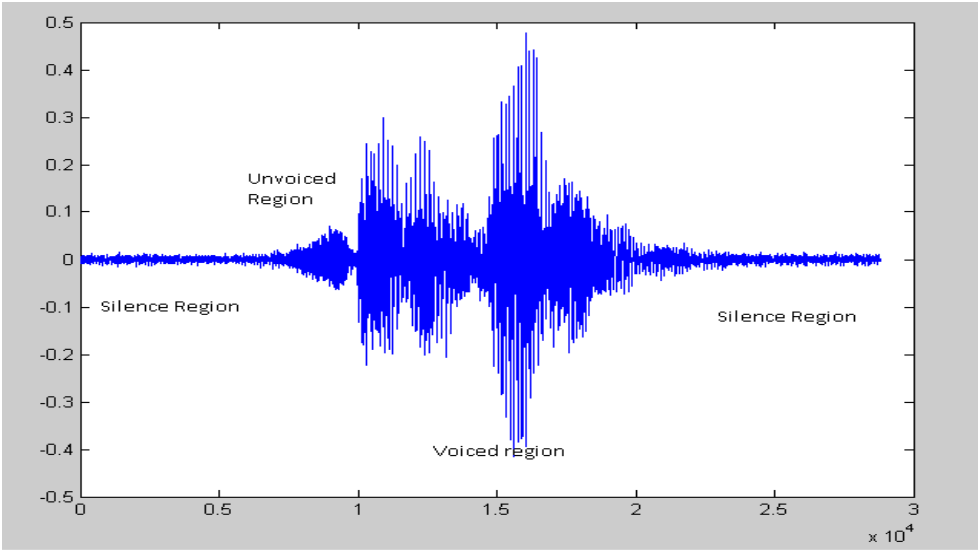

Figure 2: Sample vs amplitude waveform of word "shalgam" showing its different regions.

1.1.2 Noise

In recent years, many kinds of speech recognition systems have been proposed. However, most of the works recognize clean speech collected in quiet environments. For practical use it is necessary for recognition systems to be robust for interfering noise. Noise interference

masks the speech signal and reduces its intelligibility (Fujimoto and Ariki, 2000). Interference noise can come from acoustical sources such as ventilation equipment, traffic, crowds and commonly, reverberation and echoes. If the sound system has unusually large peaks in its frequency response, the speech signal can even end up masking itself. Many algorithms like Parallel Model Combination, Spectral Subtraction and many other kind of filtration techniques like kalman filter *etc*. have been proposed to make the interference of noise as low as possible and making the signal to noise ratio as large as possible, but still automatic speech recognition is hampered by the challenges that exist in the areas of robustness to speaker, channel and specially background noise.

1.1.3 Voice Quality

The term voice quality is used to describe the quality of sound produced with a particular setting of the vocal folds. Among numerous types of voice quality, the ones most frequently utilized across languages are modal, creaky, and breathy voice. The modal voice has well defined pitch and its fundamental frequency is relatively larger. Breathy phonation on the other side is characterized by vocal chords that are fairly abducted and have little longitudinal tension. The abduction and lowered tension allows turbulent flow of air through the glottis so turbulent noise is present across the frequency range and thus the characteristics of the breathy voice are somewhat similar to noise accept for the fact that short time energy during that phase is larger. Creaky phonation is typically associated with vocal folds that are tightly adducted but open enough along a portion of their length to allow for voicing. Due to the tight adduction, the creaky voice typically reveals slow and irregular vocal pulses in the spectrogram and also comparatively lower fundamental frequency. Voice quality distinctions are used in some languages to encode lexical contrast, and/or there may be allophonic variation in voice quality for some sounds. Voice quality also functions to signal the speaker's emotional or attitudinal status (Yoon *et al.,* 2009). Also as speakers may have different styles of speaking and if we design a recognition system that is purely able to recognize the modal voice that is the voice or the quality of speech that we use in our regular speech and some speaker speaks in other voice quality than that system may fail as the characteristics of these or we can say features on the basis of which recognition is done may differ significantly. Also there are some languages in which a similar king of word has different meaning depending on the quality of voice in which the word is being spoken.

1.2 LITERATURE SURVEY

Work on speech recognition is going on from past 50-60 years. Many advances has been made and many are to be done to improve the quality of the recognition and to make the systems speaker, voice quality and noise independent. Preprocessing is a very important phase of speech recognition which serves various purposes in many speech recognition system including silence region removal, noise removal, word boundary detection, pre-emphasis, framing and windowing *etc*. Silence removal and word boundary detection are very important out of these and for doing this the speech signal has to be divided into regions to identify where something has been spoken and where not. In the next sub sections, the literature on the topics of word boundary detection and identification of silence, voiced and unvoiced regions in speech has chronologically been surveyed.

1.2.1 Literature Reviewed for the period 1975-1990

Earlier in this era a pattern recognition approach was used (Atal and Rabiner, 1976) in which parameters like zero crossing rate, energy and LPC was used and speech signal was assigned to a particular class based on minimum distance rule obtained under the assumption that the measured parameters are distributed according to the multidimensional Gaussian probability density function. The major limitation of the method was the necessity for training the algorithm on the specific set of measurements chosen, for the particular recording conditions.

Then concept of variable decision space was used in which only three features were used avoiding time consuming linear predictive analysis but still training of the algorithm was required (Sarma and Venugopal, 1978).

A delta modulation technique was proposed (Un and Lee, 1980) for voiced, unvoiced and silence classification, in which a decision algorithm was designed based on the results of counting bit alterations of the bit stream from linear data modulation of the speech signal. When speech is coded by Linear Delta Modulation (LDM), the bit alternation rate of the LDM output bit stream gives a measure of variation of speech characteristics. The bandpass filtered decoded signal exhibits similar characteristics. Hence, with appropriate decision criteria and threshold settings, it is possible to decide whether the given segment of speech is voiced, unvoiced, or silent.

In the similar year, a robust algorithm for making the voiced-unvoiced-silence decision was proposed. This algorithm was based on a nonparametric rank-order statistical signal-detection scheme that does not require training set of data and maintains a constant false

alarm rate for a broad class of noise inputs (Benjamin and Timothy, 1980). Two rank-order decision procedures were investigated, the Kruskal-Wallis and the multiple use of the two-sample Savage statistic. The performances of these detectors were evaluated and compared to that obtained from manually classifying twenty recorded utterances. In limited testing, the average probability of misclassification of voiced speech for the average case was less than 6, 13, 28, and 55 percent, corresponding to signal-to-noise ratios of 30, 20, 10, and 0 dB, respectively.

After that cepstral subtraction techniques came into focus which was applied iteratively on the signal (Chung *et al.*, 1985). These repetitive applications of the spectral or cepstral subtraction techniques during each iteration the updates and computes a noise threshold function representing an estimation of the components of the broadband noise contained in the input signal. A scaled version of the noise threshold function is subtracted at each iteration from a corresponding function containing the noisy input signal. The subtraction provides a new signal containing less noise power than the previous signal. Experiments have shown that such iterative noise reduction methods can significantly improve the S/N of the input signal without significantly distorting the speech. Word boundary detection or the parts where something is being spoken takes place on the resultant improved signal. A significant improvement was obtained in the detection of the spoken word as well as in the recognition but still the results was not as were desired.

1.2.2 Literature Reviewed for the period 1991-2000

In this era neural networks came into the field of research and development. As neural networks are trained to do the desired work a fast training algorithm for feedforward neural nets was designed and applied to a 2-layer neural network to classify segments of speech as voiced, unvoiced and silence (Ghiselli-Crippa and Jaroudi, 1991). The speech classification method was based on features computed for each speech segment and were used as input to the network. The network weights were trained using a new fast training algorithm which uses a quasi-Newton error minimization method with a positive definite approximation of the Hessian matrix. The results indicate satisfactory performance, with percent errors in the range 3 - 5%, based on manual classification of the speech frames. But the disadvantage was that this was purely dependent on the training and performance was largely affected by the size of the training data and also was speaker dependent.

After that Voiced-unvoiced-silence classification of speech was made using a multilayer feedforward network. The network was evaluated and compared to a maximum-likelihood classifier. Results indicated that the network performance was not significantly affected by the size of training set and a classification rate as high as 96% was obtained. The feature vector for the classification is a combination of cepstral coefficients and waveform features (Qi and Hunt, 1993). The cepstral coefficients are an equivalent representation of log linear predictive (LP) spectrum of speech and provide the necessary spectral information for the classification. Additional waveform features are included to enhance the separation in pattern space when spectral information alone is not sufficient for making the classification. Six speakers (3 men and 3 women) provided speech samples for evaluating the performance of the network. The speech samples included 10 three-digit numbers and the rainbow paragraph which begins with "when the sunlight strikes raindrops in the air, it acts like a prism . . . ." Recordings were made in a quiet office environment. The results were 96% accurate but the training time was very large and also the designed network can only and only be used for the purpose of classification.

In the similar year a mapping neural network which combines unsupervised and supervised training was described and its application to the classification of segments of speech to Voiced, Unvoiced and Silence (V-UV-S) was done through computer simulations. The authors (Kia and Coghill, 1993) proposed a mapping neural network for binary inputs called the Extended Differentiator Network (EDN). EDN was used to learn the mappings from the feature vector space to the desired output vectors. The network worked well and results were faster as compared to previous neural networks and almost 90% accurate.

All these methods were performing well but only in certain conditions of no noise and only for the trained samples. Nevertheless, endpoint detection can be affected by different types of background noise over which we have no control. An explicit detector that aims to resolve the problems created by two types of noise: quasi-stationary background noise and the noise generated by the speaker was purposed (Taboada *et al*.,1994). The system was based on very simple measurements of energy, zeros crossing rate and band crossings, and was implemented on a personal computer. The real endpoint detecting efficacy of the system was tested against a set of 92 words in Spanish and the accuracy was almost 97%. But in this algorithm loss of part of the initial phoneme and in some cases of final phoneme was a major problem due to low energy at that time.

Then pitch frequency was used as a method of detection of only the voiced and the silence region of the speech means to detect the boundaries of the word. The pitch frequency (Fo)

was found to rise in a word and fall to the next word. The presence of this fall was proposed as a means of detecting word boundaries (Raman Rao and Srichand, 1996). Four major Indian languages were used and the results show that nearly 85% of the word boundaries were correctly detected. The same method used for German language showed that nearly 65% of the word boundaries were correctly detected.

A wavelet transform technique based on generating a certain mathematical function derived from the wavelet parameters that can keep track with the energy changes along the speech duration was also proposed to detect the problem of loss of begin or end of word in low-energy phonemes (Kader and Refat, 1999). The correlation model was generated from the correlation of wavelet coefficients. This model was used to generate a logical series to extract the speech duration from the whole sample and reject the noise duration at the boundaries of words. An evaluation of the system was made by superimposing a normal distributed noise to a speech signal with different signal to noise ratios. Moreover, the system was tested in the hard cases of end points such as weak fricatives at the beginning or end. The system gave a high accuracy for end point detection in normal case but not in case of low noise to signal ratio and also unvoiced region was not detected.

Another method for detecting the voiced portion of the speech or the detection of the word in noisy environment using mel scale was given after that. In this an adaptive time-frequency parameter was used for extracting both the time and frequency features of noisy speech signals. Based on the this parameter, a new word boundary detection algorithm was proposed

(Wu and Lin, 2000) by using a neural fuzzy network (called SONFIN) for identifying islands of word signals in noisy environment. Due to the self-learning ability of SONFIN, the proposed algorithm avoids the need of empirically determining thresholds and ambiguous rules in normal word boundary detection algorithms. It reduced the recognition error rate due to endpoint detection to about 10% compared to an average of approximately 50% in noisy environment.

1.2.3 Literature Reviewed for the period 2001-2012

In previous years detection of voiced parts was done either is the environment of complete silence or considering a fixed level of noise or fixed type of noise.

A word detection algorithm was proposed that could work in the presence of variable-level background noise and was tested for cars. To solve this problem, a minimum mel-scale frequency band (MiMSB) parameter was proposed which can estimate the varying

background noise level in cars by adaptively choosing one band with minimum energy from the mel-scale frequency bank (Lin *et al*., 2002). With the MiMSB parameter, some preset thresholds used to find the boundary of word signal were no longer fixed in all the recording intervals. These thresholds will be tuned according to the MiMSB parameter. An enhanced time–frequency (ETF) parameter was also proposed by extending the time–frequency parameter from single band to multiband spectrum analysis, where the frequency bands help to make the distinction between speech signal and noise. The ETF parameter can extract useful frequency information by choosing some bands of the mel-scale frequency bank. Based on the MiMSB and ETF parameters, finally a new robust algorithm was proposed for word boundary detection in variable noise-level environment. The new algorithm has been tested over a variety of noise conditions in cars and has been found to perform well not only under variable background noise level condition, but also under fixed background noise level condition. The new robust algorithm using the MiMSB and ETF gave an average accuracy of only 59% and noisy environments became the main concern of the research.

A new algorithm based on entropy feature of the speech was proposed (Weaver *et al*., 2003). This computation of the entropy estimate was carried out directly in the time domain. The original incoming speech data was first preprocessed and low frequency components were removed using a band pass filter. Then the speech signal was divided into frames of 25 ms. Then entropy profile for each frame was calculated and compared against the threshold to detect whether something is being spoken or not. The new entropy-based algorithm gave better performance in monophonic noisy environments with small to medium size vocabulary but in large size vocabulary the performance became slow.

Until now the maximum algorithms were using either zero crossing rate or energy for the classification in which results were not greater than 65% or were using neural networks in which results were very high but a lot of training was required. A new algorithm using uni-dimensional Mahalanobis Distance function was proposed for this classification was proposed. This algorithm used statistical properties of background noise as well as physiological aspect of speech production and does not assume any ad hoc threshold (Saha *et al*.,2005). In this algorithm it was checked for each sample that whether one-dimensional Mahalanobis distance function *i.e.* $|x-\mu|/\sigma$ greater than 3 or not. If it was greater than 3 sample is to be treated as voiced sample otherwise it is an silence/unvoiced. The algorithm showed almost 83% of accuracy when tested against lock number combination speech but disadvantage was that if the noise was not guassian in nature the method was failed.

As neural networks were having the limitation of the need of large amount of data for training a novel voiced-unvoiced-silence classification based on unsupervised learning was also proposed during this period (Deng and OShaughnessy, 2007). The class-dependent statistics (feature means, covariance matrices, and occurrence frequencies of voiced, unvoiced, and silence classes) needed for the classification were estimated directly from the signal to be classified via Gaussian mixture models and the expectation maximization algorithm. The classification was evaluated, and the results were encouraging as voiced, unvoiced and silence classification accuracy was greater than 91.15%.But these classification and distribution based methods were heavily relying on the distribution or first few thousand samples of the signal which are assumed to be part of noise and also can change with the passage of time.

Outlier-detection based strategies were introduced to solve this problem (Keerio *et al.* 2008). Data points, ξ_k , in a data set, ξ , that do not agree with our expectations based on the bulk of the data are termed as outliers. The popular automatic outlier-detection approaches depend on two estimates: an estimate of a nominal reference value for the data set, and a scatter estimate of the data. Based on these estimators, outliers can be detected based on the criteria

$$|\xi_k - \xi_{kr}| > \alpha\gamma \Rightarrow \xi_k = \xi_{ko}. \qquad \forall\, \xi_k \in \xi$$

The outlier-detection strategy based on the '3 σ edit rule considers the mean of the data values of the data set as the nominal reference value and the corresponding standard deviation as an estimate of the scatter. Method was not very much successful. In another strategy Based on Hampel Identifier outlier resistant median (breakpoint value of 50%) and the median absolute deviation from the median scale estimates replace the outlier sensitive mean and standard deviation estimates respectively. The method gave approximately 70% accuracy removing all the constraints but still not what was desired.

In Hindi language a careful study of the intonation pattern of language is required as there are several parameters of speech signal such as pitch, F0 fundamental frequency, duration, intensity, and pause, which can play important role in finding some clues to detect the voiced region of the speech in this language. An algorithm based mainly on two parameters namely pitches and intensity was proposed (Agarwal *et al.*, 2010). In this threshold values for pitch and intensity were calculated and compared for each frame of the speech sample. Total of 3 speakers including one male and two female speakers recorded 40 sentences and the test were done and the results were approximately 80%.

Based on this survey it can be concluded that there is not much work done of speech recognition in the Hindi language and also all the methods or algorithms that has been proposed are having one or the other limitations. Maximum algorithms studied were using zero crossing rate and energy and mel spectrum for the classification for which computation was fast but results were not very good and were not greater than 80%.On the other hand neural networks if used were giving very good results of more than 95% but they were required very large amount of data for training and were working for certain amount of vocabulary only. Also for noisy environments there is not much work being done and the accuracy is also very less.

CHAPTER 2

DATA COLLECTION, PREPROCESSING AND COMPUTATION OF FEATURES

Data collection, preprocessing and computation of features are three very important phases required before recognition phase in continuous speech recognition. In the detection of silence, voiced and unvoiced region also before applying the main algorithm for detection data collection, preprocessing (not always necessary) and computation of essential features has to be done. Preprocessing is required in the situation if data is very much affected by the noise. Also in case features are to be detected for small segments of speech signal than windowing of the signal has to be done. Section 2.1 concentrates on data collection, Section 2.2 includes preprocessing and Section 2.3 discusses algorithms to compute features. . It may be noted that these three processing stages region detection are not dependent on each other and should be planned together. Since the performance of the methods used in these stages affect the overall classification rate of the algorithm.

2.1 DATA COLLECTION PHASE

Speech in purely physical terms is nothing but a longitudinal wave which is transmitted as a pressure variation between low and high pressure with the rate of pressure variation from low, to high, to low again, determining the frequency. The degree of pressure variation (namely the difference between the high and the low) determines the amplitude. So speech is normal and best handled when stored as a vector of samples, with each individual value being a double-precision floating point number. A sampled sound can be completely specified by the sequence of these numbers plus one other item of information: the sample rate.

Recording sound in colea which is used as a tool for recording for this particular work requires the user to specify sampling frequency which provides the estimate that how many samples per second should be recorded, the duration of the recording, number of bits that should be used to store the sample value which can be either 8 or 16 bits depending on the requirement of storing the sample as floating point number or integer number and the

filename. For this particular work recordings were done at the sampling frequency of 22050 with 16 number of bits used to store the sample value and the recordings were saved in wave file format. If we plot this speech than a graph of number of samples vs amplitude is obtained which is shown in Figure 3.

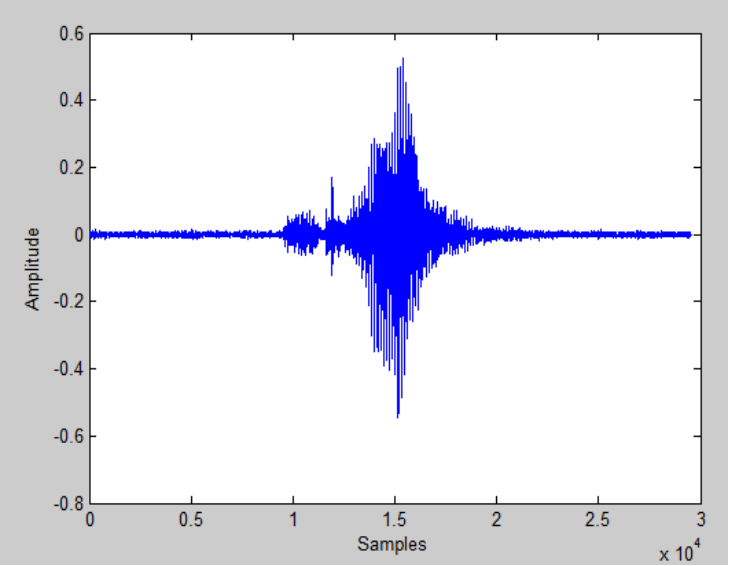

Figure 3: Typical waveform of "ghar" in Hindi (Home) plotted against number of samples and amplitude.

For this particular application the data was gathered or sounds were recorded of 3 males and one female member. 10 words were spoken 3 times by each member in Hindi like ghar (Home), bahar (outside), kabutar (pigeon), shalgam (turnip) *etc*. using laptop inbuilt microphone in mild noisy conditions.

2.2 PREPROCESSING

Preprocessing of speech signal serves various purposes in any speech processing application. It includes noise removal, pre-emphasis, windowing, framing *etc*. Only windowing and framing was needed for classification of speech as recording was done either in silence or very little background noise.

2.2.1 Windowing and Framing

Speech signal is highly variable in nature such that when examined over a short duration of time it shows similar characteristics but when examined over long duration it shows significant changes. That's why rather than working on the whole signal of speech it is being divided into frames and the process is termed ad segmentation which is the basic necessity of any audio processing system. But in segmentation it might happen that the feature is split into two: half appears in one audio frame, and the other half in another frame. The complete feature does not appear in any analysis window, and may have effectively been hidden.

To overcome this problem segmentation is done with overlap to minimize the distortion and the process is termed as windowing. The process is shown in figure 4 (Gupta, 2006). The time for which the signal is considered for processing is called is called a window and the data acquired in the window is called frame.

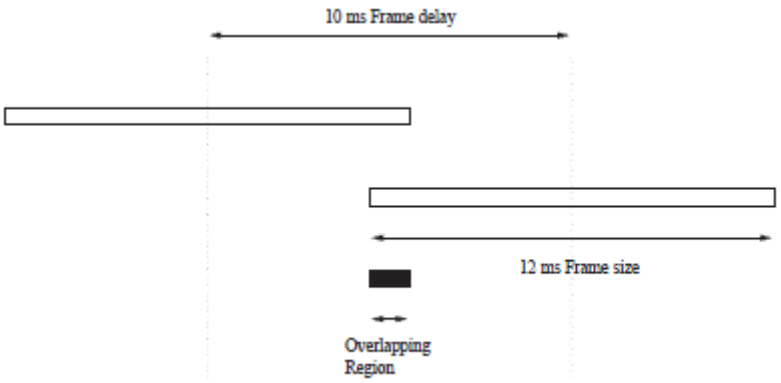

Figure 4: Windowing of speech signal showing overlap.

There are different types of windows which are used for speech processing. Some of these are rectangular window, Bartlett window and Hamming window *etc.*

For windowing a mathematical function is used which is zero-valued outside of some chosen interval. Rectangular window is the simplest kind of window and replaces all but N values of data sequences by zero where N is the size of data samples in the window. It has excellent resolution characteristics for signals of comparable strength but distortion at the boundaries is very high in this if the signal is highly variable in nature as illustrated in Figure 5.

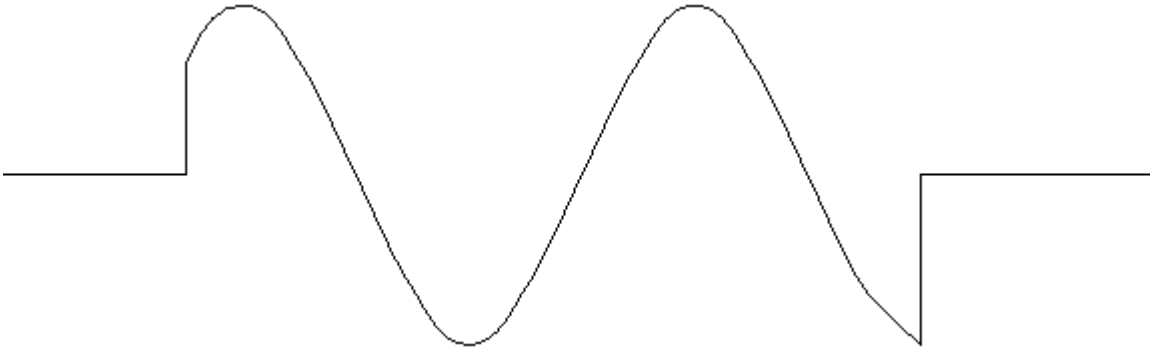

Figure 5: Distortion in rectangular window.

The most widely used window for speech processing is hamming window as it introduces least amount of distortion. This is simply a raised cosine. A typical hamming window is shown in Figure 6.

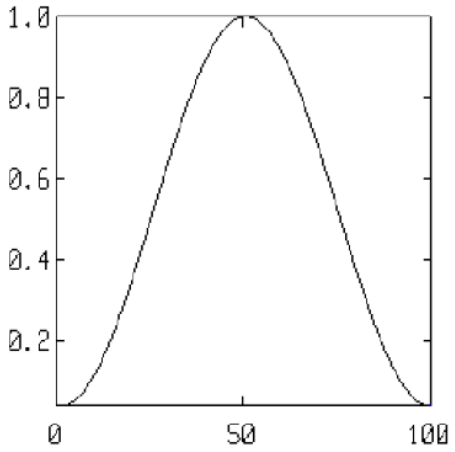

Figure 6: Impulse response of hamming window.

For this application for the short term analysis of features hamming window was used.

2.3 FEATURE EXTRACTION

Humans have the capacity to differentiate between different types of sounds. If we want a machine to identify the type of sound it will need some features or characteristics of speech on the basis of which it can differentiate different kind of regions in the speech signal to identify whether its voiced, unvoiced or silence.

In feature extraction, speech is converted into a stream of feature vectors which contain only that information about the given utterance that is important for the correct classification. These features are than stored in efficient sets of feature vectors.

Features can be classified into two basic categories.

One are the temporal features which are easy to extract, simple and have easy physical interpretation like average energy, zero crossing rate, maximum amplitude, and maximum energy *etc*. These features are generally used during preprocessing and classification of speech signal like for silence removal *etc.*

Another classification of features can be termed as spectral features. These features give quite a lot of information about the spoken phone that is why these are used for recognition of speech. For extracting these features firstly domain data is being converted into frequency domain by applying Fourier transform on it and then spectral information or spectral features are extracted from this. MECC, LPC, power spectral analyses are the different features that come under this category.

For this particular application of classification of speech signal into silence, voiced and unvoiced region, three main features are extracted after dividing the speech signal into frames using hamming window and feature vector were made which were than compared against the parameters or threshold values set for the classification.

The three different features used are the following.

Zero Crossing Rate

Short Time Energy

Fundamental Frequency

2.3.1 Zero Crossing Rate

Zero crossing rate is a measure of number of times in a given time interval or frame that the amplitude of speech signal passes through a value of zero. The rate at which zero crossing occurs is a simple measure of the frequency content of the signal. This feature is very useful for analysis and segmentation of the speech signal. This is well suited for the analysis purpose as it is virtually independent of talker volume, and apparently less speaker dependent than the spectral information or spectral features of the data.

In mathematical terms short time zero crossing rate can be defined as the weighted average of the number of times the speech signal changes sign within the time window (Rabiner and Schafer, 2007).

$$Z_n = \sum_{m=-\infty}^{\infty} 0.5|sgn\{x[m]\} - sgn\{x[m-1]\}|w[\hat{n} - m]$$

Where

$$sgn\{x\} = \begin{cases} 1 \; x \geq 0 \\ -1 \; x < 0 \end{cases}$$

Since $0.5|sgn\{x[m]\} - sgn\{x[m-1]\}|$ is equal to 1 if $x[m]$ and $x[m-1]$ have different algebraic signs and 0 if they have the same sign, it follows that $Z_{\hat{n}}$ in is a weighted sum of all the instances of alternating sign that fall within the support region of the shifted window $w[\hat{n} - m]$.

For this application rate at which zero crossing occurs was calculated by taking a window of 20 ms and a feature vector was made of the same size as that of the data sample. For example if a speech data was recorded for 2 second and as the sampling rate was 22050 than a vector dimensional array of size 44100 was made which was used in the algorithm for classification of speech signal.

A typical Speech Signal of word “kabutar” is shown in Figure 7 and its Zero Crossing Rate is shown in Figure 8.

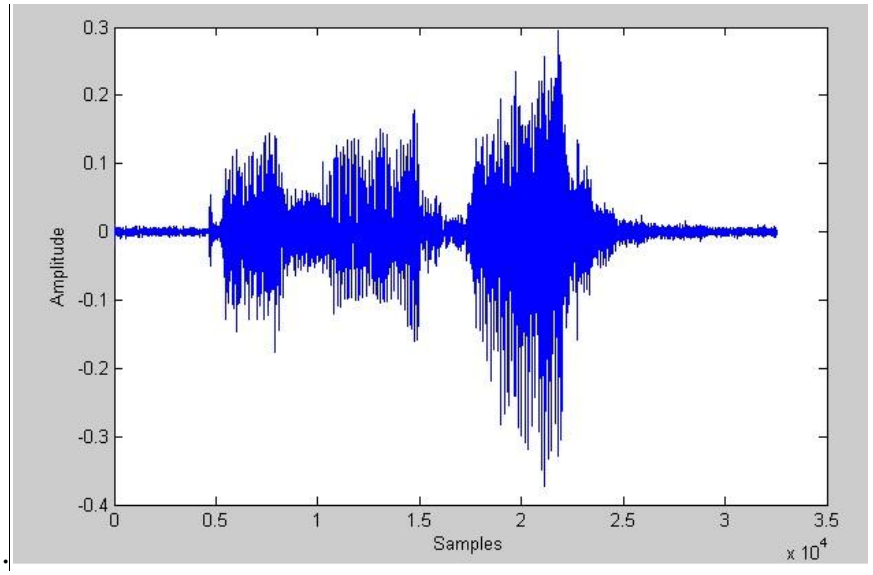

Figure 7: Speech signal of word “kabutar”.

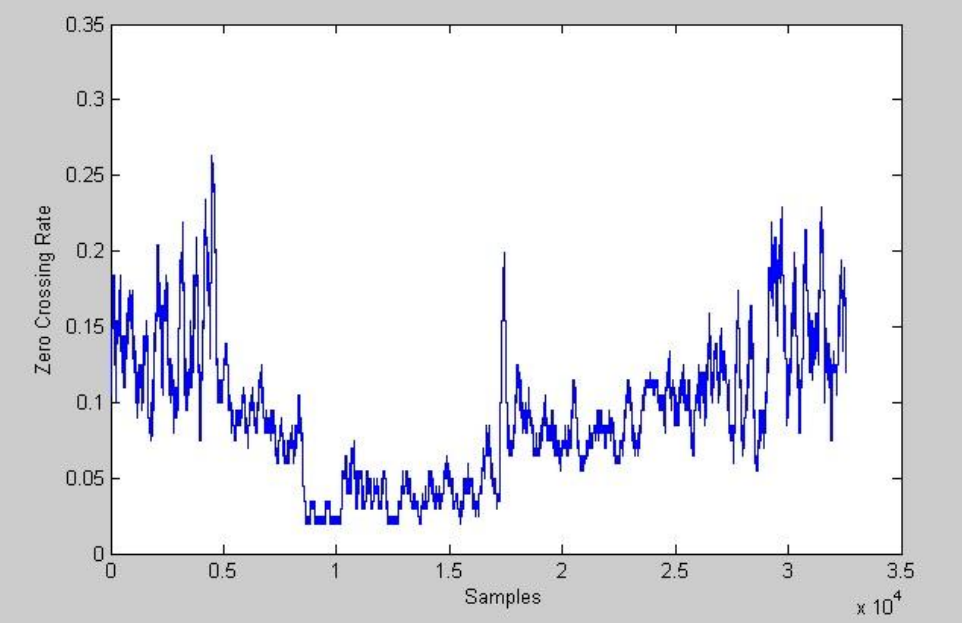

Figure 8: Zero crossing rate of word "kabutar".

2.3.2 Short Time Energy

The amplitude of speech signal varies over time. The energy of the speech signal provides a representation that reflects these amplitude variations.

For a discrete time signal x[n], the short time energy measured at sample n is defined as

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

i.e. summation of squared amplitude in an N-sample frame ending at n.

But as speech signal is highly variable in nature and assumed to have stationary properties only within a short time frame so short time energy is also calculated after doing windowing of the signal and if windowing is used short time energy is defined as

$$E_n = \sum_{m=-\infty}^{\infty} (x[m]w[n-m])^2$$

Where n lies between zero and N-1 where N is the size of the window. If N is very large E_n would change very little and will not reflect the variations in the amplitude. This effect is illustrated in Figure 10.

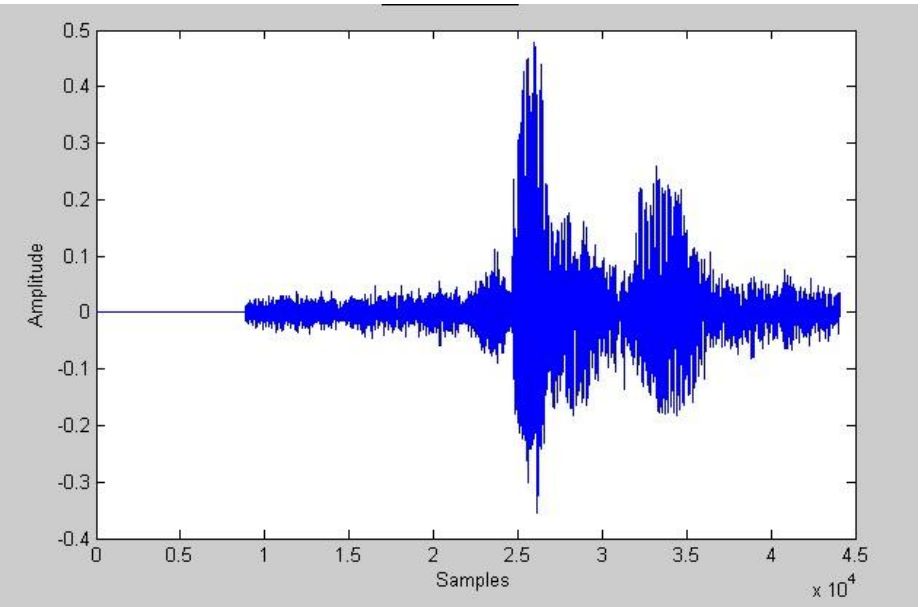

Figure 9: Speech signal of word “shalgam”.

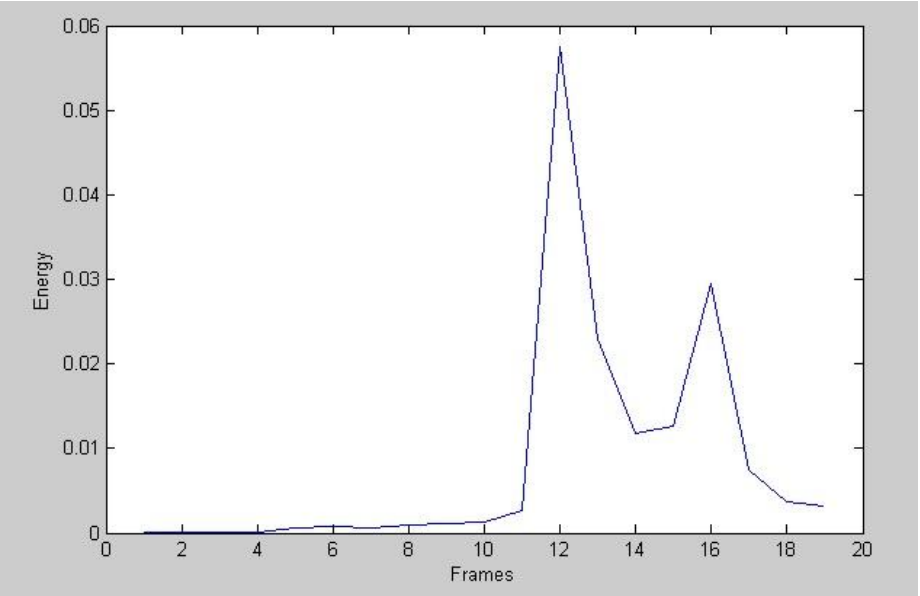

Figure 10: Effect of taking large window on short time energy.

On the other hand if N is too small energy function would not be a smooth one. This effect is shown in Figure 11.

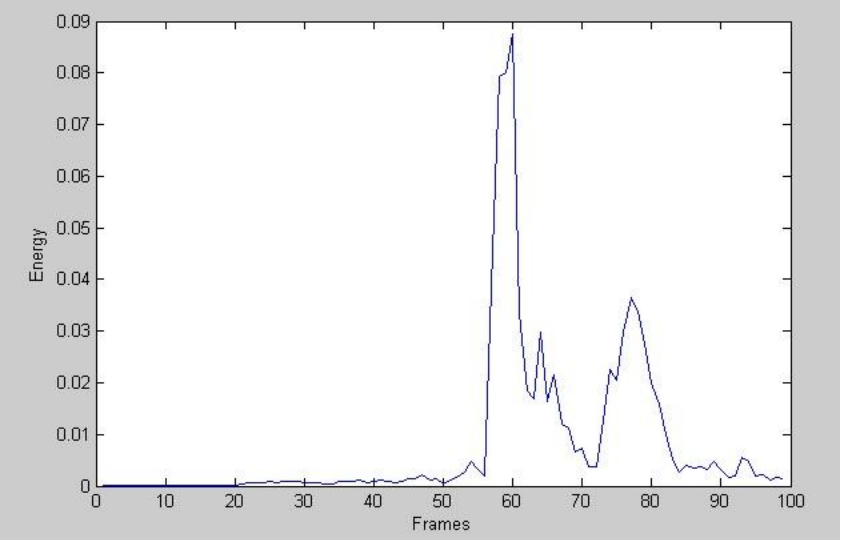

Figure 11: Effect of taking small window.

So for this particular work hamming window of size 50 ms was taken and according to the threshold value the classification was done. It is illustrated in Figure 12.

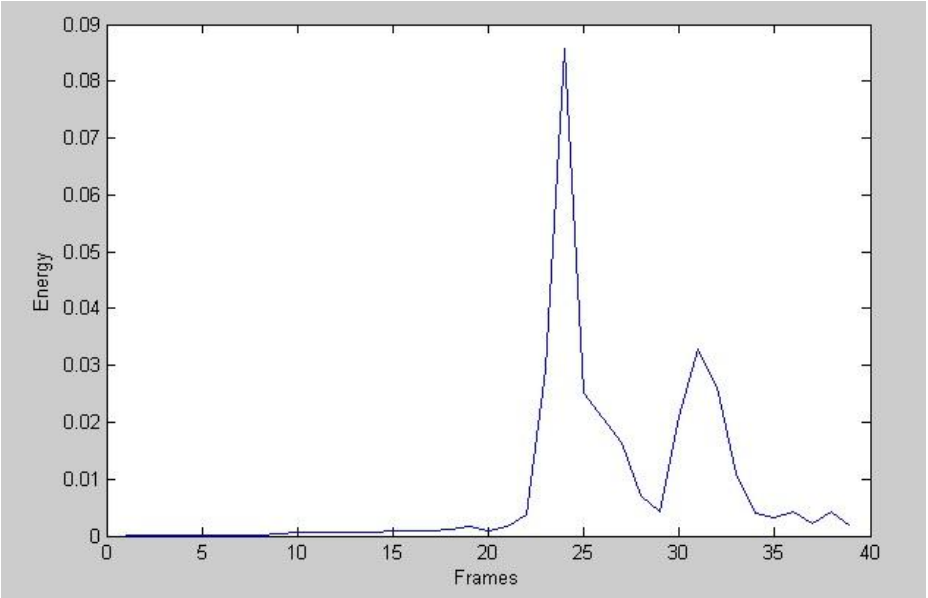

Figure 12: Short time energy of word "shalgam" taking average window of 50 ms.

For the classification purpose some threshold value needs to be set. If the threshold value was set to be static for example in this particular case it was observed for many words that the energy in the voiced region was greater than 30 db but it could not be taken into consideration if the words were spoken in different intensity or loudness. So threshold value was calculated dynamically according to the speech data (http://www.mathworks.com/matlabcentral/fileexchange/28826-silence-removal-in-speech-

signals). For calculating the threshold value after getting the feature vector of the energy following process was carried out:

- Compute the histogram of the feature sequence's values.
- Detect the histogram's local maxima.
- Let M1 and M2 be the positions of the first and second local maxima respectively. The threshold value is computed using the following equation:

$$T = \frac{W.M_1 + M_2}{W + 1}$$

W is a user-defined parameter. Large values of W obviously lead to threshold values closer to M1. Short time energy of "shalgam" with threshold indicated by black line in shown in Figure 13.

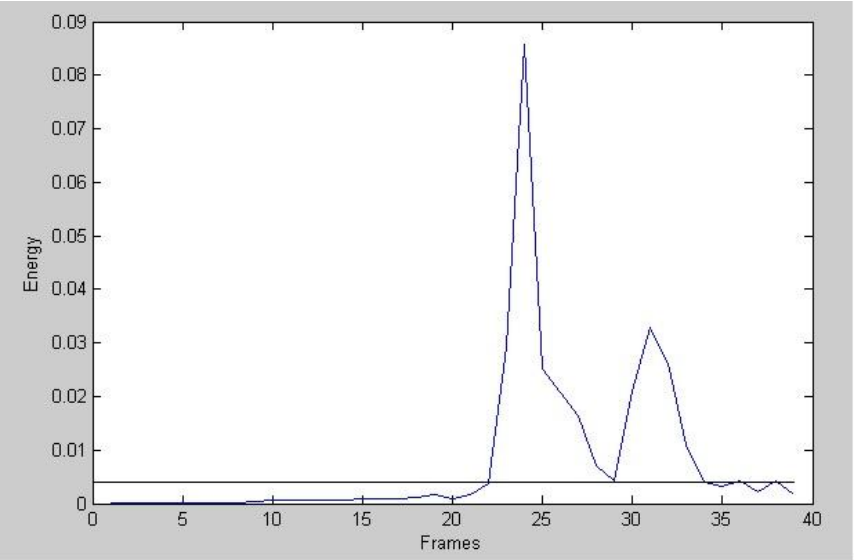

Figure 13: Short time energy of "shalgam" showing threshold value.

On the basis of this dynamic threshold calculated with other feature values the classification of speech signal was done.

2.3.3 Fundamental Frequency

Fundamental frequency also known as pitch is usually the lowest frequency component, or partial, which relates well to most of the other partials. In a periodic waveform, most partials are harmonically related, meaning that the frequency of most of the partials is related to the frequency of the lowest partial by a small whole-number ratio. The frequency of this lowest

partial is the fundamental frequency of the waveform and how this fundamental frequency changes over the period of time is determined by the f0 contour.

There are various time domain and frequency domain approaches available for calculating the fundamental frequency of the speech signal.

Time domain approach could use some time-related features such as ZCR, peak picking, and autocorrelation. From this approach it is easy to calculate the pitch as computation complexity is less but accuracy is not better than frequency domain analysis. Out of these autocorrelation is one of the most robust and reliable method of pitch detection. Autocorrelation preserves information about harmonic and formant amplitudes in speech signals by measuring the similarity of the signal and its time delay (Zhao *et al*., 2007). Effect of autocorrelation approach is shown in Figure 15.

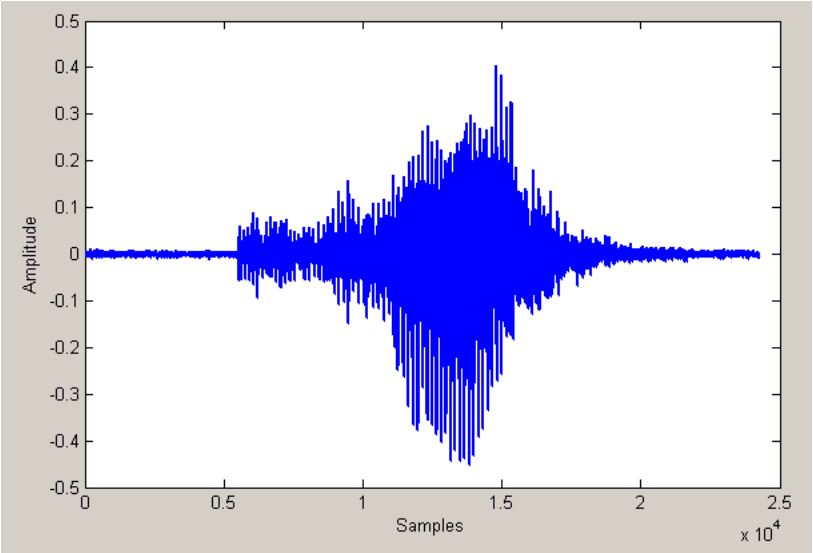

Figure 14: Speech signal of word “ghar”.

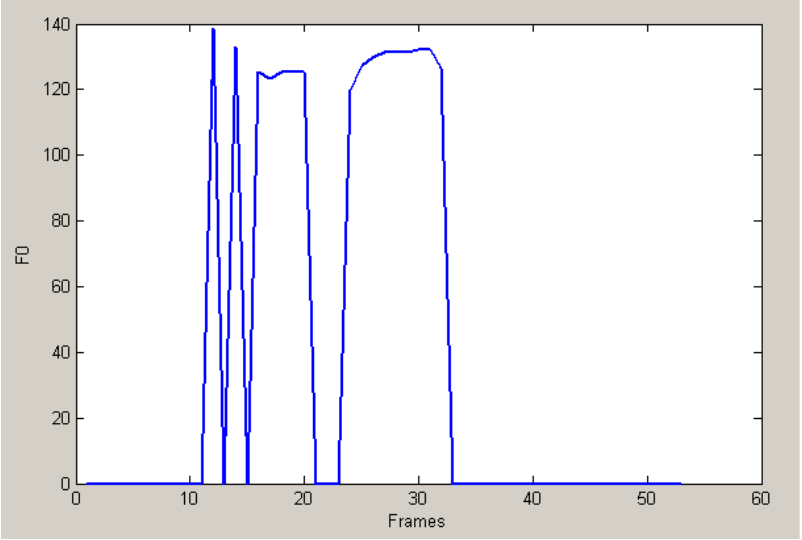

Figure 15: F0 using autocorrelation approach.

Frequency domain analysis could apply, for example, to cepstrum and harmonic matching. These approaches generally have higher accuracy than time domain methods. A reliable way of obtaining an estimate of the dominant fundamental frequency for long, clean, stationary speech signals is to use the cepstrum, and the cepstrum pitch detector performed much better on lower pitch speakers than on higher pitch speakers. The cepstrum is a Fourier analysis of the logarithmic amplitude spectrum of the signal. If the log amplitude spectrum contains many regularly spaced harmonics, the Fourier analysis of the spectrum will show a peak corresponding to the spacing between the harmonics: *i.e.*, the fundamental frequency. Effectively, we are treating the signal spectrum as another signal, and then looking for periodicity in the spectrum itself.

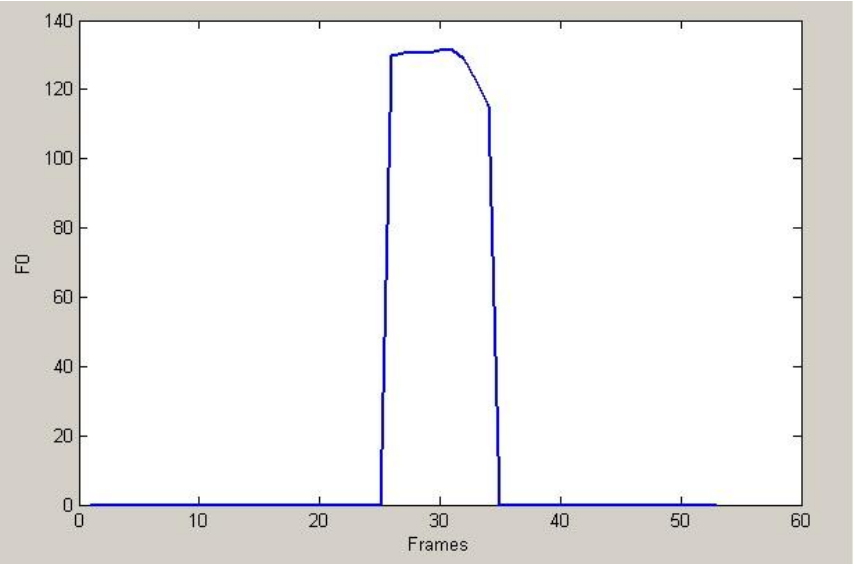

Figure 16: F0 using cepstrum.

For this work cepstrum approach was used and the fundamental frequency was calculated for frames of 40 ms. convey very different meanings with variations in intonation. However, accurate pitch detection is difficult partly because tracked pitch contours are not ideal smooth curves. So a smoothing algorithm for detected pitch contours was also applied after calculating the normal fundamental frequency.

The pitch frequency (F0) is found to rise in a word and fall to the next word (Rao and Shrichand, 1996). This characteristic of fundamental frequency is used for this application to classify silence, voiced and unvoiced region of speech.

CHAPTER 3

IDENTIFICATION OF VOICED, UNVOICED AND SILENCE CHUNKS IN SPEECH SIGNAL

The classification of speech signal into voiced, unvoiced provides a preliminary acoustic segmentation for speech processing applications, such as speech synthesis, speech enhancement, and speech recognition where speech synthesis refers to the artificial production of human speech and speech recognition is the conversion of speech into text.

Voiced Speech is produced when periodic pulses of air generated by the vibrating glottis resonate through the vocal tract. Vowels which are the most interesting class of sound in any language and on which most practical speech recognition systems rely to achieve high performance are the parts of voiced region of speech signal. Other than this the group of semivowels and nasal consonants like /m/, /n/ also come under this category. About two-thirds of speech is voiced and this type of speech is also what is most important for intelligibility.

Unvoiced speech is non-periodic, random-like sounds, caused by air passing through a narrow constriction of the vocal tract. Fricatives like /f/, /sh/, /s/ which are produced by exciting the vocal tract by a steady air flow, becomes turbulent in the region of a constriction in the vocal tract come under the category of unvoiced sounds. The frequency during the pronunciation of unvoiced sounds is very high and hence the zero crossing rate also is high (Rabiner and Juang, 1993).

Silence speech on the other hand is having least energy of all the parts of the speech signal.

For this work of classification the characteristics of features calculated above were noticed for different kind of regions and the results were extracted from them as which are described in section 3.1 of this chapter. These results serve as the basis for making the algorithm for the classification which is described in detail in section 3.2 of this chapter.

3.1 RESULTS AND FACTS OBSERVED FROM FEATURES

On the basis of studies of the three calculated features on various words the facts that came over were satisfactory and useful for the classification of voiced, unvoiced and silence regions in input speech signal.

3.1.1 Facts observed from zero crossing rate

Voiced speech is produced as a result of excitation of approximately 12 db thereby producing a concentration of energy at low frequencies; this is reason why voiced speech usually shows a low zero crossing count. The zero-crossing count for silence can vary considerably from one speaking environment to another reflecting the variable characteristics of the room noise. But generally in quite conditions the zero-crossing count for silence is expected to be lower than for unvoiced speech, but larger or comparable to that for voiced speech (Atal and Rabiner, 1976).

For the voiced region of the speech signal it was observed that zero crossing rate was always less than 0.1 and for the silence region it was between 0.1 and 0.3. It is shown in Figure 17.

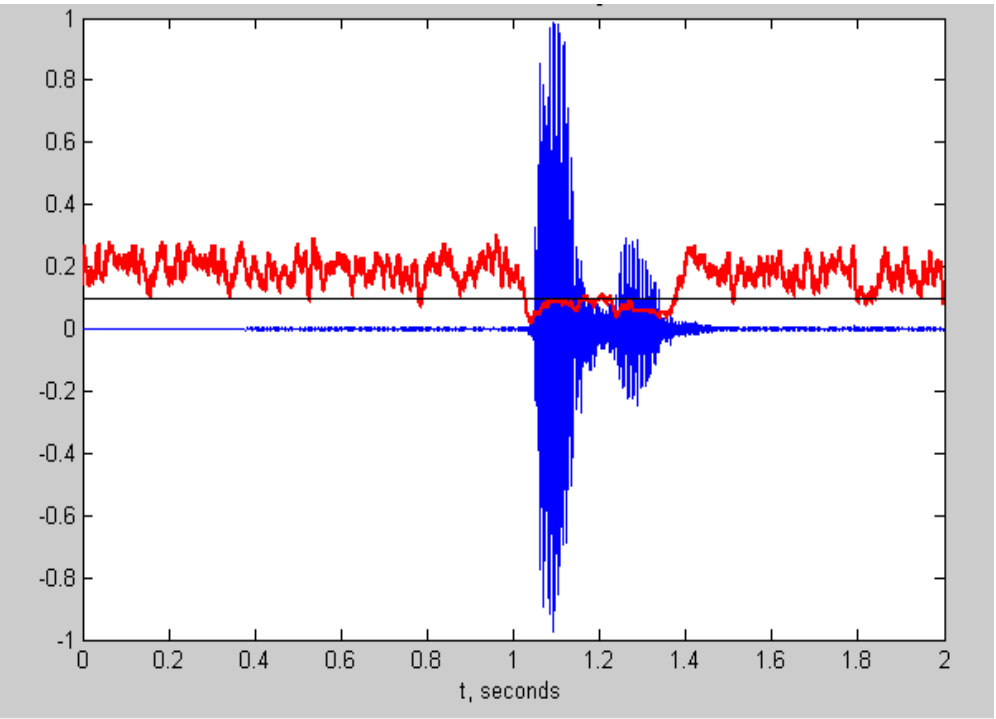

Figure 17: Zero crossing rate plotted over signal for word “bahar” .

The red line in the above fig shows the zero crossing rate and blue line is showing the signal. The black line separation is clearly indicating that the value of zero crossing rate when the word is spoken is less than 0.1 and when there is silence it is between 0.1 and 0.3 with a little error in the voiced region where for very less samples the zero crossing rate is becoming greater than 0.1 and for silence region it is becoming less than 0.1.

Unvoiced speech is produced due to excitation of the vocal tract by a noise-like source at a point of constriction in the interior of the vocal tract. While the spectrum of the noise source is flat, the vocal-tract response usually increases with frequency. Thus, the unvoiced speech has a concentration of energy at high frequencies and shows a high zero crossing count (Atal and Rabiner, 1976).

For the unvoiced region or the region where the fricative or unvoiced sound was produces it was noticed that zero crossing rate was always greater than 0.3.It is shown in Figure 18.

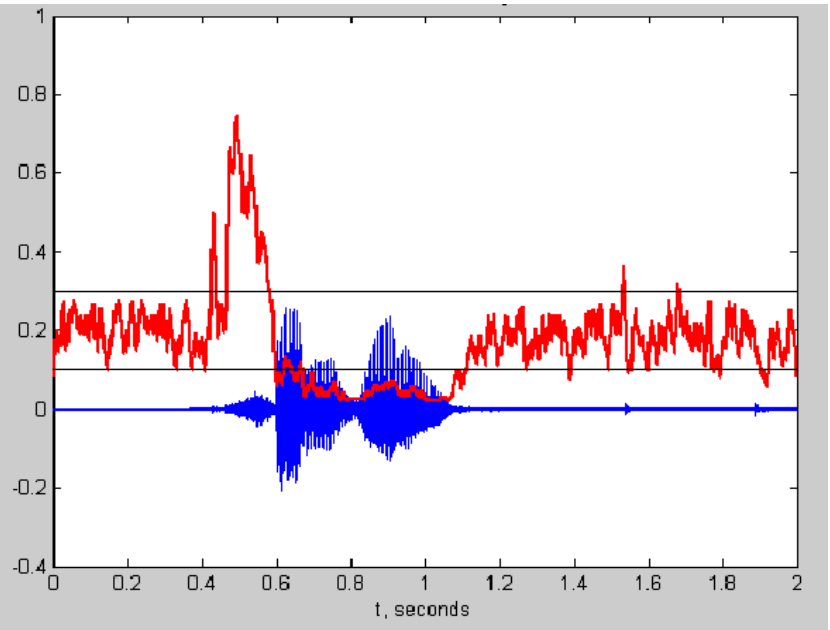

Figure 18: Zero crossing rate plotted over signal for word “samajhdaar”.

So this characteristic could be successfully used for the classification of voiced, unvoiced and silence region in the input speech signal.

But when something breathy was spoken which results in turbulent flow of air through the glottis, it was observed that the zero crossing rate was similar to that of the silence or little noisy region .In Figure 19 word “bahar” was spoken in breathy voice and it was noticed that the zero crossing rate was coming between 0.1 and 0.3 which was the threshold value set for

the classification of the silence region. So in this case zero crossing rate could not be used alone for the classification.

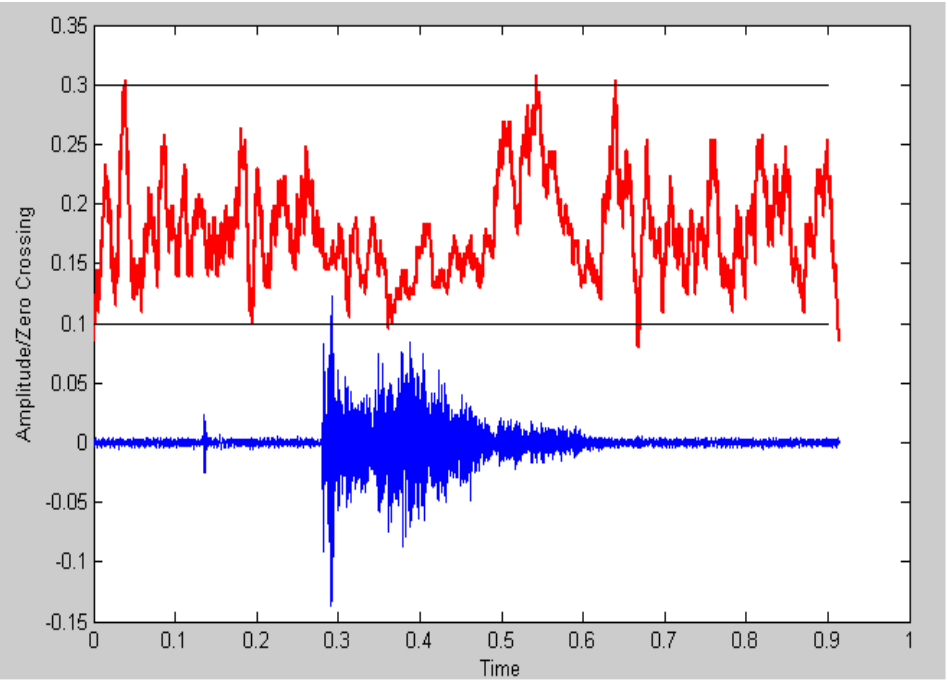

Figure 19: Signal and ZCR of word "bahar".

Also sometime silence region was having less zero crossing rate due to some constant background sound as shown in the Figure 20.

Also unvoiced region was having sometimes less zero crossing rate. So it was concluded that zero crossing rate alone cannot be used for classification.

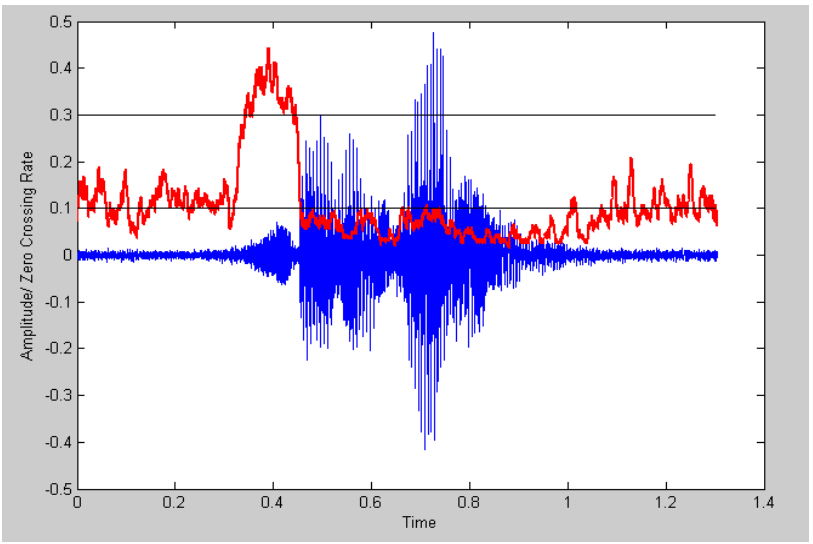

Figure 20: Signal and ZCR for word "shalgam" showing ZCR less than 0.1 in silence region.

3.1.2 Facts Observed from Short Time Energy

The energy of the voiced sound is much higher than the energy of silence and the energy of unvoiced sounds is lower than for voiced sounds, but often higher than for silence (Sarma and Venugopal, 1978).

For voiced region always the short time energy was found to be greater than dynamically calculated threshold value as shown in the following fig where the black line shows the threshold value of the short time energy of the signal and it can be seen that where there is no silence or some pause in the word than energy is always greater than the threshold value. This is being shown in Figure 21.

So the problem of having errors sometimes due to zero crossing rate could be easily solved using short time energy for the purpose of classification.

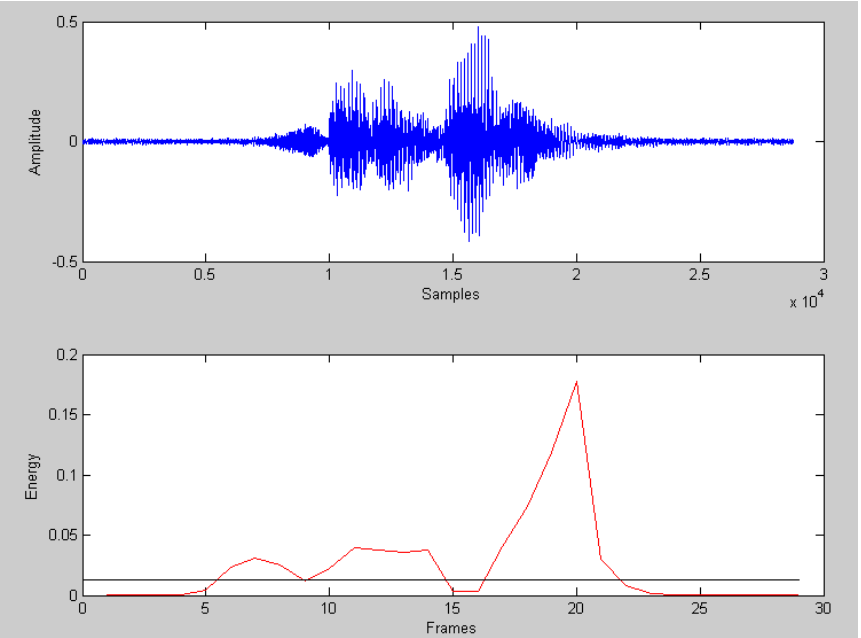

Figure 21: Signal and STE of "kabutar".

Also the difficulty of breathy voice having zero crossing rate similar to that of silence region was also solved by taking short time energy into consideration as short time energy of this kind of voice was always greater than threshold. It is being shown in Figure 22.

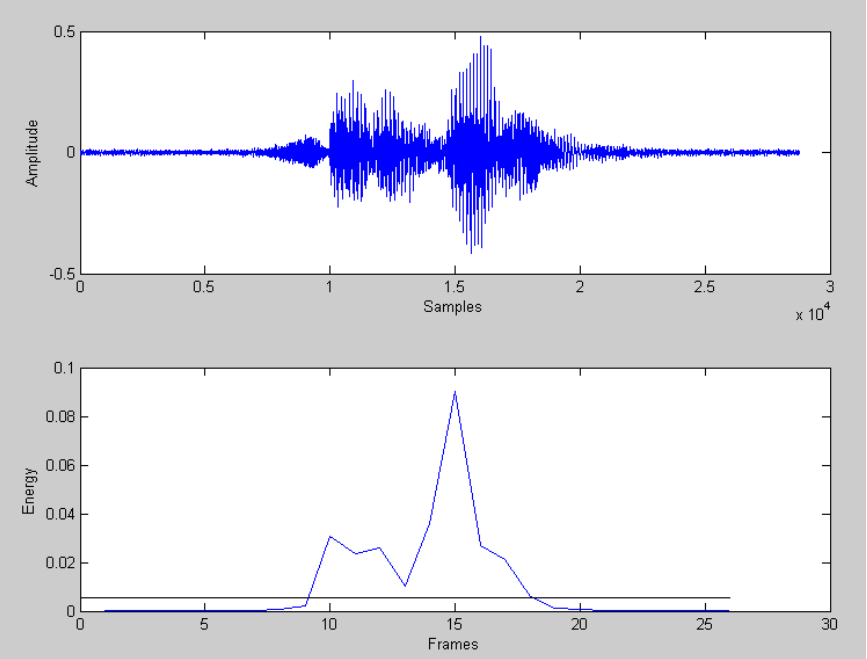

Figure 22: Signal and STE of word "shalgam" .

3.1.3 Facts observed from fundamental frequency

It is the quality of pitch that it raises when something is spoken voiced and then falls. So for the unvoiced and silence region fundamental frequency is always zero as shown in Figure 23.

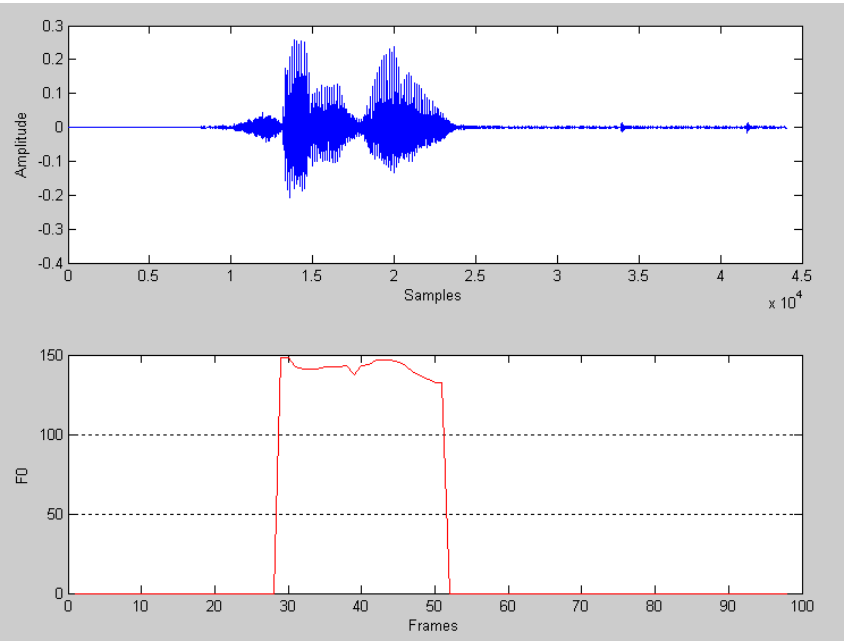

Figure 23: Signal and F0 of "shalgam" showing zero value of F0 in unvoiced region.

So the problem with the zero crossing rate of sometimes not having the zero crossing rate very large or very near to the upper limit of the voiced region this characteristic of fundamental frequency could be successfully used.

3.2 ALGORITHM USED FOR CLASSIFICATION

After analyzing the results from different features calculated the algorithm was designed for the identification of silence, unvoiced and voiced chunks in speech signal. This algorithm was then implemented in MATLAB 2011a. The steps used for identification and flow chart are explained in the subsequent sections.

3.2.1 Algorithm

Step 1: Read the sound file and create speech vector SV(n) where n is the length of speech signal.

Step 2: Take 20 ms rectangular window and calculate ZCR vector (ZC) of the same length as that of speech signal.

Step 3: Compute Fundamental Frequency Vector (F0) taking hamming window of size 40 ms.

Step 4: Map F0 vector to the vector of length similar to that of speech signal.

Step 5: Compute Short Time Energy vector (STE) taking Hamming window of 50 ms.

Step 6: Map STE vector to the vector of length similar to that of speech signal.

Step 7: Calculate threshold T_E for STE.

Step 8: Make output vector (OUT) of length equal to SV and initialize all its values to zero.

Step 9: Repeat for i=1 to n
    If ZC(i) < 0.1 and STE(i) > T_E then set OUT(i)=0.1(for voiced)
    Else if ZC(i) between 0.1 and 0.3 and STE(i) < T_E
        If F0(i)=0 then set OUT(i)=0.2(for silence)
        End if
    Else if ZC(i) > 0.3
        If F0(i)=0 then set OUT(i)=0.3(for unvoiced)
        End if
    End if

Step 10: plot SV and OUT.

3.2.1 Flow Chart

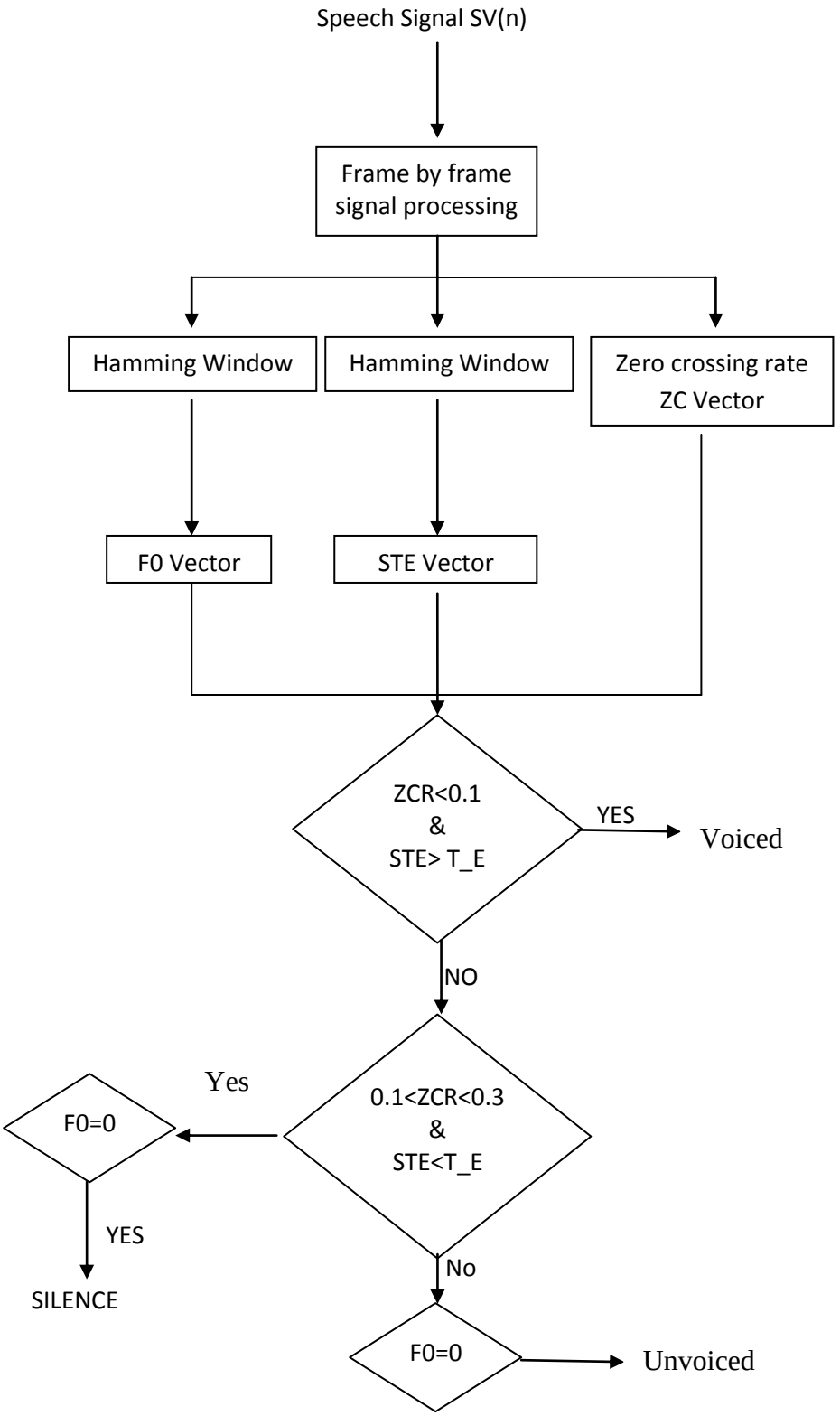

Figure 24: Flow chart of the algorithm.

The above described algorithm was applied to all the collected data or the words in Hindi which were spoken by four different speakers. Comparing all the feature vectors with the threshold values set for the automation, a final OUT matrix was made. In this matrix a value of 0.1 was given to the samples of sound that were voiced in nature, a value of 0.2 was given to the samples that were in silence and 0.3 value was given to the samples were fricative or unvoiced sound was produced. Finally this OUT matrix was plotted over signal.

The detailed results and conclusions are discussed in the next chapter.

CHAPTER 4

RESULTS AND DISCUSSION

After applying the algorithm discussed in the previous chapter the output that was in the form of a matrix, was of the same length as that of the length of the speech signal. For example if a word of 2 second was spoken at a sampling frequency of 22050 Hz than a matrix of length 44100 was formed indicating the state (voiced, unvoiced or silence) of the speech at each sample point.

In the section 4.1 the outputs obtained when the algorithm was applied to different words is shown and in section 4.2 accuracy of the algorithm is discussed.

4.1 OUTPUTS OF ALGORITHM

The simplest output of the algorithm is shown in Figure 25. In this word “bahar” is spoken by a male and 2 second speech signal is being recorded and passed in the algorithm as input.

In this output is divided into three equal parts explaining each and every step of the algorithm.

In the first part of the output speech signal is plotted in sample vs. amplitude form and is plot in blue color. Then ZCR is calculated and is plotted over this signal which is shown in green color.

In the second part of the output F0 of the signal is calculated dividing speech signal into 98 frames of equal length and is plotted against samples. In this it can be clearly shown that when nothing is being spoken F0 is zero otherwise not.

In the last section similar to F0, STE of the signal is calculated dividing it into frames and then it is plotted, clearly indication that when nothing is spoken STE is less than threshold.

Finally the calculated output matrix is plotted over signal in red color to indicate the voiced and silent part of the speech.

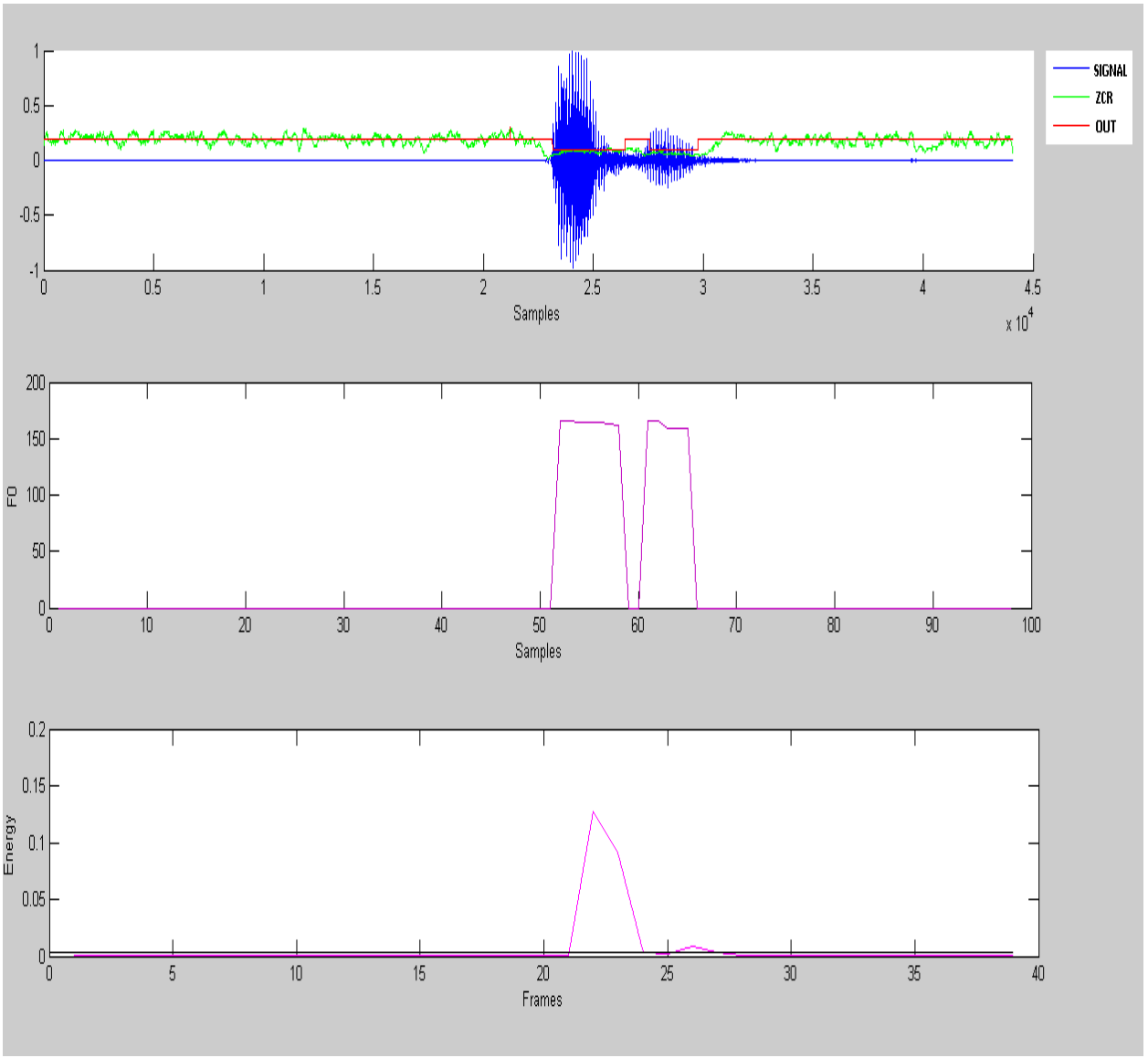

Figure 25: Output of algorithm for word “bahar”.

The above example is identifying all the parts almost correct and with very high degree of accuracy. Similarly when “kabutar” was spoken by a male speaker than also a very high degree of accuracy as is shown in Figure 26. In this Figure only the signal and the output matrix is plotted ti indicate the voiced and silence region in the speech.

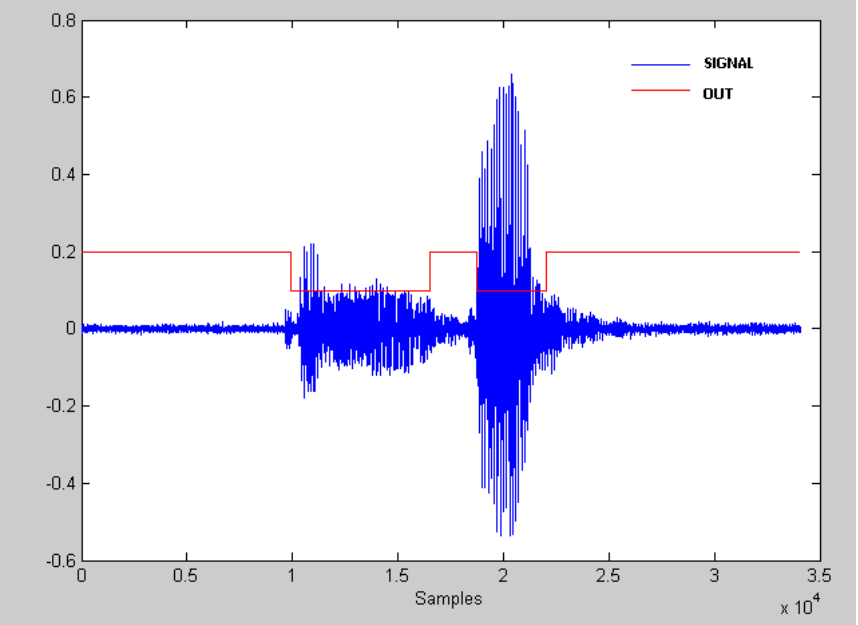

Figure 26. Output of algorithm for word "kabutar".

But there were some cases when the algorithm was not able correctly able to correctly identify very few samples of speech. For example in Figure 27 word "shor" was spoken by a female speaker and it can be clearly seen that some of the unvoiced region (were fricative /sh/ was spoken) came into the category of silence and some of the silence region (in the end) came into category of unvoiced due to noise increasing the ZCR of the signal.

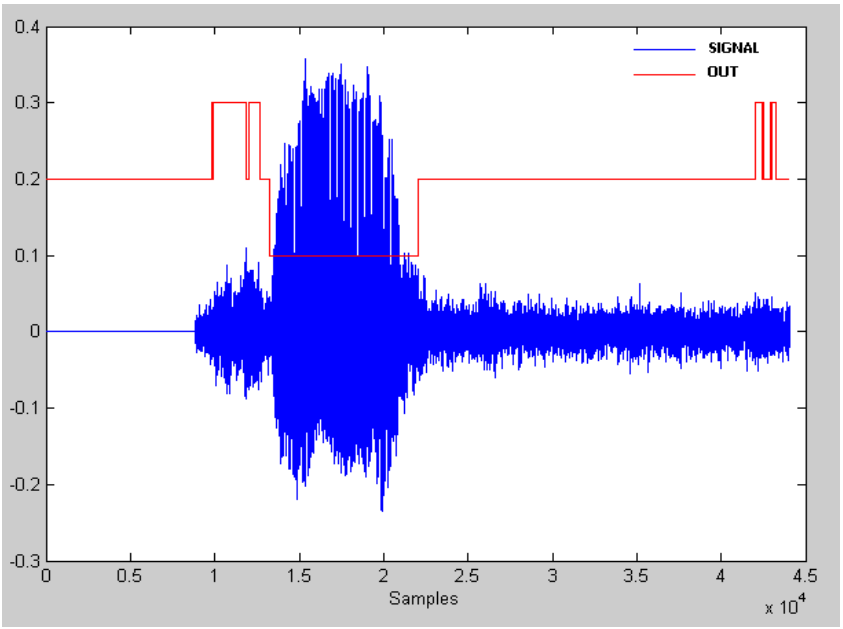

Figure 27. Output of algorithm for word "shor" spoken by female speaker.

Similar case for the word “shalgam” when spoken by a male member can be seen in Figure 28 indicating incorrect identification for very small number of samples in the unvoiced region (in the starting) and in the end where there was some noise created while speaking.

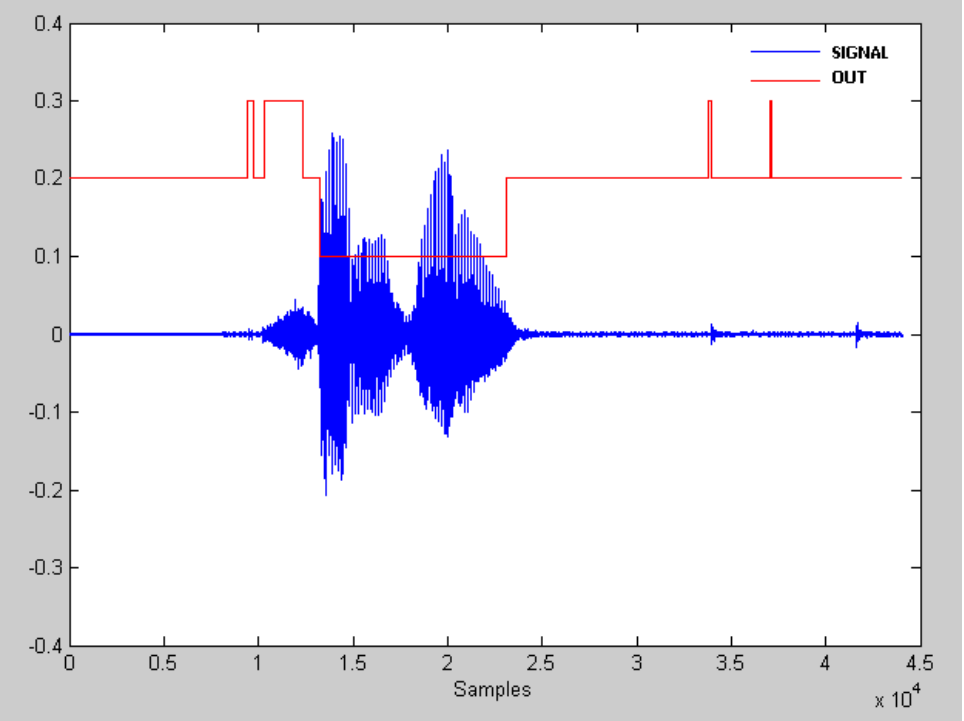

Figure 28. Output of algorithm for word “shalgam” spoken by male speaker.

4.2 ACCURACY OF ALGORITHM

When this algorithm was tested for the complete data set a good accuracy was found. Accuracy of the algorithm was calculated by checking how many samples in the spoken word were identified correctly compared to the manual classification of the voiced, unvoiced and silence region in the word and then dividing them by total number of samples.

The accuracy of the algorithm for four different speakers taking all the 15 words spoken 3 times is shown in Table 1, 2, 3 and 4.

The overall accuracy of the algorithm is shown in Table 5.

Table 1: Accuracy of first speaker (male)

Word Spoken	**Accuracy (Spoken 1st time)**	**Accuracy (Spoken 2nd time)**	**Accuracy (Spoken 3rd time)**	**Average accuracy**
"ajay"	99.73	99.71	99.2	99.54
"kabutar"	95.7	98.6	97.5	97.26
"shalgam"	92.6	90.9	98.9	94.13
"ghar"	94.6	94.7	93.6	94.3
"bahar"	91.4	89.85	90.8	90.68
"aag"	95	94.3	96.6	95.3
"aam"	98.4	96.1	97.27	97.25
"dhaga"	96.8	98.18	93.75	96.24
"gadi"	98.9	99.3	95.4	97.86
"ghadi"	98.43	96.45	96.78	97.22
"ghas"	97	99.5	99.18	98.56
"hawa"	97.2	98.13	97.8	97.71
"kal"	98.7	99.2	99.1	99
"mitti"	95.4	94.13	92.1	93.87
"shor"	97.89	98.54	98.2	98.21
Average accuracy for speaker 1				96.47

Table 2: Accuracy of second speaker (male)

Word Spoken	Accuracy (Spoken 1st time)	Accuracy (Spoken 2nd time)	Accuracy (Spoken 3rd time)	Average accuracy
"ajay"	93.9	94.3	94.2	94.13
"kabutar"	96.4	97.2	92.3	95.3
"shalgam"	95	96.2	93.6	94.93
"ghar"	92.4	86.4	95.7	91.5
"bahar"	98.4	89.9	92.3	93.53
"aag"	98.7	98.34	99.18	98.74
"aam"	98.2	98.56	96.7	97.82
"dhaga"	96.7	98.4	93.24	96.11
"gadi"	99.52	94.9	90.45	94.96
"ghadi"	96.7	95.62	96.9	96.41
"ghas"	94.42	88.2	92.3	91.64
"hawa"	96.5	97.8	98.1	97.47
"kal"	98.2	99.13	98.46	98.6
"mitti"	96.4	92.61	97.28	95.43
"shor"	92.1	86.8	88.9	89.27
Average accuracy for speaker 2				95.06

Table 1, 2 and 4 give the accuracies for the male speakers, clearly indicating that the algorithm was able to identify the different regions at a very good rate. An average accuracy of 96.03 % was found for male speakers.

Table 4 shows the accuracy of the algorithm for the female speaker. The algorithm for female speaker was showing better results and accuracies.

Table 3: Accuracy of third speaker (female)

Word Spoken	**Accuracy (Spoken 1st time)**	**Accuracy (Spoken 2nd time)**	**Accuracy (Spoken 3rd time)**	**Average accuracy**
"ajay"	99.5	98.8	98.8	99.03
"kabutar"	96.5	98.18	97.7	97.46
"shalgam"	97.75	96.59	99.5	97.95
"ghar"	97.5	97.73	98.18	97.80
"bahar"	99.2	98.6	98.8	98.87
"aag"	98.4	98.12	97.24	97.92
"aam"	99.1	98.62	99.32	99.01
"dhaga"	99.47	97.67	99.23	98.79
"gadi"	97.16	98.7	98.6	98.15
"ghadi"	98.68	99.13	99.2	99
"ghas"	97.25	97.8	97.34	97.46
"hawa"	98.7	95.67	97.54	97.3
"kal"	98.74	99.56	98.9	99.07
"mitti"	98.3	98.57	98.18	98.35
"shor"	99.6	99.23	97.34	98.72
Average accuracy for speaker 3				98.33

Table 4: Accuracy of fourth speaker (male)

Word Spoken	**Accuracy (Spoken 1st time)**	**Accuracy (Spoken 2nd time)**	**Accuracy (Spoken 3rd time)**	**Average accuracy**
"ajay"	97.29	98.64	98.8	98.24
"kabutar"	96.4	98.59	97.27	97.42
"shalgam"	95.5	97.02	95.4	95.97
"ghar"	99.45	97.25	98.6	98.43
"bahar"	96.75	98.4	98.4	97.85
"aag"	96.85	97.50	93.40	95.92
"aam"	97.08	97.42	96.05	96.85
"dhaga"	94.35	96.63	98.25	96.38
"Gadi"	94.88	94.44	94.82	94.71
"Ghadi"	95.36	97.46	95.60	96.14
"Ghas"	97.69	95.05	93.03	95.26
"Hawa"	98.02	99.24	97.09	98.12
"Kal"	94.74	97.24	94.53	95.5
"Mitti"	94.64	96.42	98.76	96.61
"Shor"	93	96.78	96.5	95.43
Average accuracy for speaker 4				96.58

Table 5: Overall accuracy of algorithm

Speaker	Average accuracy	Overall accuracy
1	96.47	**96.61**
2	95.06	
3	98.33	
4	96.58	

It was noticed that maximum errors or wrong identifications were coming in the starting and ending of the word. But still the accuracy of algorithm was good showing 96.61% results.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

Speech Recognition is in research for many years. After years of research and development the accuracy of Automatic Speech Recognition still remains one of the important research challenges. The design of Recognition systems requires careful attention to the issues like speech representation, preprocessing, feature extraction *etc.*

The classification of speech as voiced, unvoiced and silence is one of the most fundamental and difficult problem encountered in speech processing and if it is detected accurately it can increase the efficiency and accuracy of the recognition system to a great extent.

This thesis was also an attempt to develop an algorithm that could solve this problem. Three fundamental features namely: ZCR, STE and F0 was used in the algorithm for the classification purpose and an accuracy of 96.61 % was achieved. The errors in the system were mainly in the starting and the ending of the word due to little noise or lower energy during the starting and ending of the word.

The results achieved in present study motivate to extend the present work to achieve a higher degree of accuracy. Also the algorithm is showing less accuracy in very noisy conditions. So still improvements are needed in the algorithm to make it robust and helpful in achieving a high degree of accuracy for recognition.

REFERENCES

1) Agarwal, A., Jain, A. and Prakash, N., 2010. Word Boundary Detection in Continuous Speech based on Suprasegmental Features for Hindi Language. 2nd International Conference on Signal Processing Systems, vol. 2, pp. V2-591-V2-594.
2) Anusuya, M. A. and Katti, S. K., 2009. Speech Recognition by machine: A review. International journal of Computer science and information security, vol. 6, no. 3, pp. 181-205.
3) Atal, B. S. and Rabiner, L. R., 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to Speech Recognition. IEEE transactions on acoustic, speech, and signal processing, vol. 24, no. 3, pp. 201-212.
4) Bachu, R.G., Kopparthi, S., Adapa, B., Barkana, B. D., Separation of Voiced and Unvoiced using Zero-Crossing Rate and Energy of the Speech Signal. American Society for Engineering Education (ASEE) Zone Conference Proceedings, 2008.
5) Chung, M., Kushner, W. M. and Damoulakis, J. N., 1985. Word Boundary Detection and Speech Recognition of Noisy Speech by Means of Iterative Noise Cancellation Techniques. IEEE International Conference on ICASSP. Vol. 10, pp. 1838.
6) Cox, V. B. and Timothy, L. V., 1980. Nonparametric rank order statistics applied to robust voiced-unvoiced-silence classification. IEEE transactions on acoustic, speech, and signal processing, vol. 28, issue. 5, pp. 550-561.
7) Deng, H. and O'Shaughnessy, 2007. Voiced-Unvoiced-Silence speech sound classification based on unsupervised learning. IEEE international conference on Multimedia and Expo, pp. 176-179.
8) Fujimoto, M. and Ariki, Y., 2000. Noisy speech recognition using noise reduction method based on Kalman filter. IEEE transactions on acoustic, speech, and signal processing, vol. 3, pp. 1727-1730.
9) Ghiselli-Crippa, T. and El-Jaroudi, A., 1991. A fast neural net training algorithm and its application to voiced-unvoiced-silence classification of speech. IEEE transactions on acoustic, speech, and signal processing, vol. 1, pp. 141-144.
10) Gupta, R, 2006. Speech Recognition for Hindi, M. Tech Thesis, Dept. Computer Science and Eng. IIT Bombay, Bombay.

11) Kadel, N. S. A. and Refat, A. M., 1999. End point detection for noisy speech using a wavelet based algorithm. Proceedings of sixth national conference on Radio Science, pp. C18/1-C18/5.
12) Keerio, A., Mitra, B. K., Birch, P., Young, R., and Chatwin, C., 2008. On Preprocessing of Speech Signals. International Journal of Signal Processing, Vol. 5, no. 3, pp. 216-222.
13) Kia, S. J. and Coghill, G. G., 1993. A Mapping Neural Network and its Application to Voiced-Unvoiced-Silence Classification. First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, pp. 104-108.
14) Lin, C., Lin, J. and Wu, G, 2002. A robust word boundary detection algorithm for variable noise-level environment in cars. IEEE transactions on intelligent transportation systems, vol. 3, no. 1, pp. 89-101.
15) McLoughlin, I., 2009. Applied Speech and Audio Processing. Cambridge University Press.
16) Qi, Y. and Hunt, B. R., 1993. Voiced-Unvoiced-Silence Classifications of Speech Using Hybrid Features and a Network Classifier. IEEE transactions on speech and audio processing, vol. 1, no. 2, pp. 250-255.
17) Rabiner, L. and Juang, B. H., 1993. Fundamental of Speech Recognition. PTR Prentice-Hall, New Jersey.
18) Rabiner. L. and Schafer, R. W., 2007. Introduction to Digital Speech Processing (Foundations and trends in Signal Processing). Now Publications, Netherlands.
19) Raman Rao, G. V. and Srichand, J., 1996. Word boundary detection using pitch variations. Forth international conference of spoken language, vol. 2, pp. 813-816.
20) Saha, G., Chakroborty, I. and Senapati, S. 2005. A new silence removal and endpoint detection algorithms for speech and speaker recognition applications. In proceedings of NCC, pp. 56-61.
21) Sarma, V. V. S. and Venugopal, D., 1978. Studies on pattern recognition approach to voiced-unvoiced-silence classification. IEEE International Conference on ICASSP, Vol. 3, pp. 1-4.
22) Taboda. J., Feijoo, S., Balsa, R. and Harnandez, C., 1994. Explicit estimation of speech boundaries. IEE proceedings of science, Measurement and Technology, vol. 141, issue. 3, pp. 153-159.

23) Un, C. K. and Lee, H. H., 1980. Voiced/unvoiced/silence discrimination of speech by Delta Modulation. IEEE transactions on acoustic, speech, and signal processing, vol. 28, no. 4, pp. 398-407.
24) Weaver, K., Waheed, K. and Salem, F. M., 2003.An Entropy based Robust Speech Boundary Detection Algorithm for Realistic Noisy Environments. International joint conference on Neural Networks, vol. 1, pp. 680-685.
25) Wu, G. and Lin, C., 2000. Word Boundary Detection with Mel-Scale Frequency Bank in Noisy Environment. IEEE transactions on signal and audio processing, vol. 8, no. 5, pp. 541-554.
26) Yoon, T.J., Zhuang, X., Cole, J. and Jhonsen, M. H., 2009. Voice Quality Dependent Speech Recognition. Linguistic patterns in spontaneous speech, Academia Sinica.
27) Zhao, X., O'Shaughnessy, D. and Minh-Quang, N., 2007. A Processing Method for Pitch Smoothing Based on Autocorrelation and Cepstral F0 Detection Approaches. International Symposium on Signals, Systems and Electronics, pp. 59-62.