Automatic Identification of Modal, Breathy and Creaky Voices

A Thesis

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

Master of Technology

Submitted by

Ajay Sharma

(Roll No. 601003001)

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 Modal, Breathy and Creaky Voices"**, 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.

(Ajay Sharma)

Roll No.: 601003001

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

8.6.12

(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 in the past few decades have changed a lot, from size of almost a room to size of one's palm. Nowadays, even mobile phones are equivalent to a mini computer. Interacting with computer has changed from punched cards to finger tip, but still speech is not widely used as an interaction medium with the computer. This is mainly due to the problems faced during recognizing of speech.

Voice quality is one of the reasons for the not so fast and effective growth in the domain speech recognition. This thesis deals with the identification of modal, creaky and breathy voices. An algorithm is presented in this thesis which successfully identifies these three types of voice qualities. The thesis is divided into five chapters. A brief outline of each chapter is given in the following paragraphs.

Chapter 1 firstly discusses the basic model of Speech Recognition. Then the issues in Automatic Speech Recognition are discussed which are: noise, voice quality and detection of voiced, unvoiced and silence region. Finally a literature survey on the algorithms and methods used to identify different types of voice qualities is done.

Chapter 2 is divided into three parts, *i.e.*, data collection, preprocessing and computation of features. Data collection part describes how data was collected and for how many users it was collected. The preprocessing phase then discusses the preprocessing technique applied (windowing) before features are extracted. Finally feature extraction explains the different features used, like zero crossing rate, fundamental frequency and short time energy.

Chapter 3 discusses the facts and results obtained from the features used which are then used to identify the different voice qualities. Finally an algorithm is designed using these features and applied to the data collected.

Chapter 4 is divided into two parts the first part displays the output obtained from the algorithm for words spoken in different voice qualities. The next part shows the accuracy obtained for different voice qualities along with the overall accuracy of the algorithm. The algorithm proposed is able to achieve 90.1% accuracy in identifying the modal voices, 89.8 accuracy for breathy and finally 80.7% for creaky voices.

Chapter 5 concludes the work. It is worth mentioning here that overall accuracy achieved in this work using the proposed algorithm is 87.2%. Also future scope in this domain is discussed in this chapter.

ACKNOWLEDGEMENTS

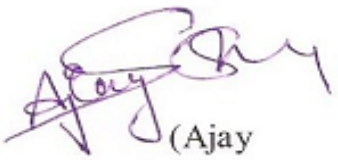
My sincere thanks to all the people around me who helped me in completing this thesis work. First, I wish to thank **Dr. R. K. Sharma** (Professor) of School of Mathematics and Computer Applications, Thapar University, Patiala for giving me an opportunity to work under his guidance. His continued support, 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, Poonam 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: 8th June 2012

Place: Thapar University, Patiala.



(Ajay Sharma)

LIST OF FIGURES AND GRAPHS

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Expanded Forms
ASR	Automatic Speech Recognition
SVM	Support Vector Machine
ECG	Electrocardiogram
HNR	Harmonic to Noise Ratio
FFT	Fast Fourier Transform
NACF	Normalized Auto Correlation Function
APP	Aperiodicity Periodicity
HOS	Higher Order Stastics
CART	Classification And Regression Tree
HMM	Hidden Markov Model
ADC	Analog to Digital Converter
DAC	Digital to Analog Converter
F0	Fundamental Frequency
ZCR	Zero Crossing Rate
DFT	Discrete Fourier Transform
STE	Short Time Energy

CONTENTS

CHAPTER 1

INTRODUCTION

Speech is one of the most important aspects of our life. Expressing one's thought without speech is impossible. In the early days keyboard and mouse were the only way to communicate with computer. As time passed and with advancement in technology human beings started using speech as a way to communicate with computer, but due to the random nature of speech it has not been very successful till now. There are many hindrances and issues such as noise, voice quality of user *etc*. that make the problem complex.

Actually, a computer or a machine is not many times expected to understand what is uttered. But it is expected to be controlled via speech or to transcript the acoustic signal to symbols. The ultimate goal of research on Automatic Speech Recognition (ASR) is to build machines that are indistinguishable from humans in the ability to communicate in natural spoken language. In this sense, speech recognition is not a mature science but an emerging one.

Speech Recognition (also known as **Automatic Speech Recognition** (ASR), or Computer 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.

1.1 Basic Model of Speech Recognition:

The basic model of a general Speech Recognition System will have the following components:

Sound Recording and word detection component, feature extraction component, speech recognition component, acoustic and language model.
The basic model of speech recognition system is shown in Figure 1.1

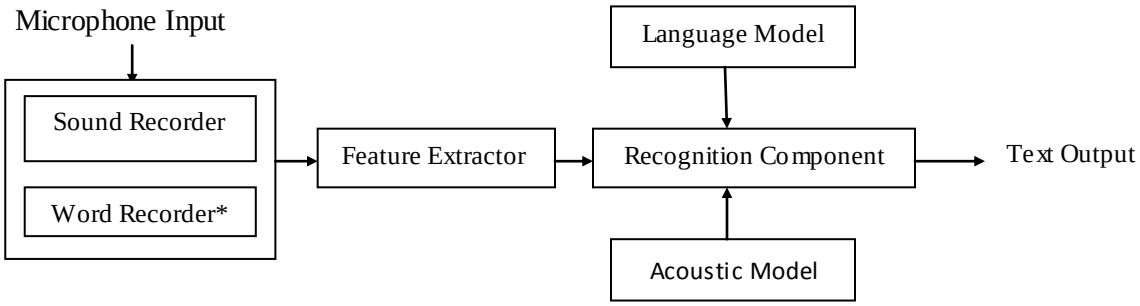


Figure 1.1: Basic model of speech recognition system

- Sound recording and word detection component: This component takes the input from microphone and identifies the presence of words. Word detection is done using energy and zero crossing rate of the signal. The output of this component can be a wave file or a direct feed for the feature extractor.
- Feature extraction component: In this component the feature vectors for the sound signals given to it are extracted. It generates Mel Frequency Cepstrum Coefficients and Normalized energy as the features that are used to uniquely identify the given sound signal.
- Recognition component: This component recognizes the sound using the features generated by the previous component. The recognizer can be a Continuous, Multi-dimensional Hidden Markov Model based component, or a SVM or Neural Net based. It is responsible for finding the best match in the knowledge base, for the incoming feature vectors.
- Knowledge model: The acoustic and language model are used by the recognition system while recognizing the. Acoustic Model has a representation of how a word sounds.

1.2 Issues in Speech Recognition

Work on Speech Recognition started in the early 1980s but even after 4 decades of work there are still many issues that are yet to be solved and are of very importance. Some of the most important issues are discussed in this section.

1.2.1 Detection of voiced, unvoiced and silence region

In the world of computer, speech is nothing but a graph plotted between time and amplitude which is a combination of voiced sounds, unvoiced sounds and silence region. Now to detect where which type of sound exists is a very important task as it will make the further detection much easier. If we can successfully detect these regions, then the techniques used in isolated word recognition can be used for continuous word recognition with some alterations.

1.2.2 Noise

Noise in case of Speech recognition means adverse environmental conditions for example people speaking in background, some machines operating causing noise, wind blowing heavily and many others. Recognition accuracy of Speech recognition Systems decreases in such conditions. One relationship between the strength of the speech signal and the masking sound is called the signal-to-noise ratio, expressed in decibels. Ideally, the S/N ratio is greater than 0 dB, indicating that the speech is louder than the noise.

1.2.3 Voice Quality

After decades of research and work many ASRs have been developed but almost all of them work under constrained environment and when exposed to other environment their recognition performance degrades. One of the main reasons for the performance degradation is voice quality. Almost none of them takes into consideration the voice quality of the speaker.

The term voice quality describes the quality of sound produced with a particular setting of the vocal folds which includes modal phonation that is produced by a plain or normal voice quality and the non-modal phonation of breathy and creaky voice qualities. If we can reliably extract acoustic features that differentiate phones that differ from each other regarding voice quality, then such a difference can be modeled in an ASR with the goal of improving recognition performance.

1.3 Literature review

Work in the domain of ASR started in the late 1970s and since then 5 decades have passed and a good amount of research has been done and many recognizers have been developed such as BYBLOS by Y.L. Chow *et al.* (1987) and SPINIX by Kai-Fu Lee *et al.* (1990), and many others. Each one of these recognizers works under specific constraints and none of these considered voice quality of speaker as a major issue. The main research in the domain of voice quality dependent recognition began in the decade of 1990.

In the starting of the 1990s, the work started on how to identify the different types of voice qualities with Childers and Lee (1991) publishing a paper on their work done to examine several factors of voice quality that might be affected by changes in vocal ford vibratory patterns. They classified voice quality into four types: modal, creaky (vocal fry), falsetto and breathy. They used glottal waveform obtained from ECG and speech features like turbulent noise and waveform peak factor to distinguish between these different types. Krom (1993) developed a cepstrum-based technique for determining Harmonics-to-Noise Ratio (HNR) in speech signals. The method involved discrimination between harmonic and noise energy in the magnitude spectrum by means of a comb-liftering operation in the cepstrum domain. Sensitivity of HNR to (a) additive noise and (b) jitter was tested with synthetic vowel-like signals generated at 10 fundamental frequencies. All jitter and noise signals were analyzed at three window lengths in order to investigate the effect of the length of the analysis frame on the estimated HNR values. His method was then used by Ratree (1995) for the Acoustic and perceptual investigation of breathy voice. They tested the algorithm on three speakers (2 male, 1 female) of Javanese producing a word list of 31 minimal breathy/clear word pairs. Results showed that the algorithm reliably distinguished breathy from clear tokens for all three speakers, with higher HNRs for clear than for breathy tokens.

Hillenbrand *et al.* (1994) evaluated the effectiveness of acoustic measures which were made of signal periodicity, first harmonic amplitude and spectral tilt. They found that Periodicity measures provided the most accurate predictions of perceived breathiness, accounting for approximately 80% of the variance in breathiness ratings. The relative amplitude of the first harmonic correlated moderately with breathiness ratings, and two measures of spectral tilt

correlated weakly with perceived breathiness. Childers and Ahn (1995) modeled features of the glottal volume-velocity waveform for three voice types: modal voice, vocal fry (creaky voice), and breathy voice. The study analyzed data measured from two sustained vowels and one sentence uttered by nine adult, male subjects who represented examples of the three voice types. The primary analysis procedure was glottal inverse filtering, which estimated the glottal volume-velocity waveform.

Hillenbran (1996) extended his previous work done in 1994 by extending the results obtained previously to speakers with laryngeal pathologies and conducted tests using connected speech in addition to sustained vowels. Breathiness ratings were obtained from a sustained vowel and a 12-word sentence spoken by 20 pathological and 5 non pathological talkers. The acoustic measures were the same for the sustained vowels, a frequency domain measure of periodicity provided the most accurate predictions of perceived breathiness, accounting for 92% of the variance in breathiness ratings. The relative amplitude of the first harmonic and two measures of spectral tilt correlated moderately with breathiness ratings. For the sentences, both signal periodicity and spectral tilt provided accurate predictions of breathiness ratings, accounting for 70%-85% of the variance. Gobl and Chasaide (1997) submitted their work which was published in the book "The Handbook of Phonetic Sciences". In their work they discussed various determinants of voice source variations, and their role in linguistic, paralinguistic strands of spoken communication. Also they discussed how to analyze the voice source like to obtain glottal flow they used inverse filtering. Gordon (1998) provided a cross-linguistic perspective of non-modal vowels. Some languages of the world have vowels characterized by non-modal phonation, *e.g.,* breathy voiced vowels, voiceless vowels or creaky vowels. His work showed that Some languages of the world have vowels characterized by non-modal phonation, *e.g.,* breathy voiced vowels, voiceless vowels, or creaky vowels and depending on the language and on the phonation type, non-modal vowels may either contrast with or be allophonic variants of modal voiced vowels. He also concluded that non-modal vowels have a quite different distribution from modal vowels. First, they are quite rare cross-linguistically, both as phonemic segments which contrast with modal vowels, and as non-contrastive allophones of modal voiced vowels. Another characteristic property of non-modal vowels which differentiates them from modal vowels is their limited distribution. For example, voiceless vowels are often limited to word-final position, and creaky vowels tend to occur adjacent to glottalized consonants. In other languages, non-modal vowels

are the synchronic manifestations of other types of contrasts, *e.g.*, segmental, tonal, or durational ones.

Gordon (2001) studied Linguistic aspects of voice quality with special reference to Athabaskan (a group languages spoken by large group of people in North America). He studied two languages Hupa and Western Apache. Finally he concluded that for these two languages if we take the FFT spectra of the speech signal and study the spectral tilt relations we can easily differentiate between modal, creaky voice and breathy voice. Until now the study of difference between modal and non-modal (breathy and creaky) voice has been spanning disciplines from linguistics to biomedicine and from physics to music appreciation. For example, linguists have examined the way in which changes in voice quality signal changes in meaning or offer clues to the grammatical structure of utterances; otolaryngologists are interested in voice quality as a symptom of disease; and singing teachers are concerned with how voice quality changes across a singer's range. The variety of questions motivating studies of non-modal phonation has resulted in a confusing literature that is spread across many journals, and that reflects the different priorities, methods, and terminological traditions of unrelated academic areas. Gerratt and Kreiman (2001) reviewed the literature on several types of non-modal phonation, and attempted to unify descriptions across disciplines and descriptive domains. They also provided an example of the kind of study which they believed is needed to establish unique, valid categories for phonation types.

Gordon (2001) extended his previous works along with Ladefoged they explained the cross-linguistic distribution of phonation contrasts, such as voiced and unvoiced contrasts in English, modal and breathy nasals in Newar and modal and creaky nasals in Kwakw'ala. The most important work done by them was finding the Phonetic properties associated with phonation types like periodicity, acoustic intensity, spectral tilt, fundamental frequency and formant frequencies. Shrivastav and Sapienza (2003) published paper on objective measures of breathy voice quality obtained using an auditory model. They used two methods to describe breathy voice subjective and objective. Subjective methods usually take the form of listener's ratings of voice quality, often made on an equal-appearing interval or visual analog scale. Objective measures, on the other hand, make specific measurements from the vocal acoustic signal or from other physiological signals associated with voice production. These measures relate to vocal fold

physiology underlying the production of breathy voices. Ten listeners rated 27 voice stimuli using a five-point rating scale. Acoustic measures were determined from these stimuli and were selected based on their history of having a moderate to strong correlation to perceptual ratings of breathiness. They obtained new measure which included the partial loudness of the signal and the loudness of the aspiration noise. Measures obtained from the output of the auditory model were found to a count for a high amount of variance in the perceptual ratings of breathiness. Ishi (2004) published his work on Analysis of Autocorrelation-based Parameters for Creaky Voice Detection. He defined a Normalized Autocorrelation Function (NACF) and using this function he proposed some NACF based parameters Peak Magnitude Ratio, Peak Position Ratio, Peak Width Ratio, Maximum peak magnitude, Maximum peak position, and Maximum peak width. Finally he concluded that the ratio between the first two autocorrelation peaks (NACR) was found to be the primary parameter to discriminate between modal and creaky phonation.

Vishnubhotla and Wilson (2005) used APP Detector (Aperiodicity/ periodicity/pitch) developed by Deshmukh and Wilson (2003). They refined the working of APP detector based on the known features of creaky voice like fundamental frequency substantially lower, relative amplitude of fundamental component reduced. Finally they tested the refined APP detector on speech files by the same male speaker, same utterance in both modal and creaky voice. The APP detector performed well and identified creaky frames as being creaky. Voiced fricative frames were also successfully separated. Ishi *et al*. (2008) proposed a method for the automatic detection of Vocal fry (creaky voice). They used acoustic parameters like power peak detection, intra-frame periodicity measure and impulse similarity measure. The basic idea of the proposed method was to scan for local power peaks in a "very short-term" power contour for obtaining glottal pulse candidates, check for periodicity properties, and evaluate a similarity measure between neighboring glottal pulse candidates for deciding the possibility of being vocal fry pulses. In the periodicity analysis, autocorrelation peak properties were taken into account for avoiding misdetection of periodicity in vocal fry segments. Evaluation of the proposed acoustic measures in the automatic detection resulted in 74% correct detection.

Malyska and Quatieri (2008) introduced a general signal-processing framework for interpreting the effects of both stochastic and deterministic aspects of non-modality on the short term spectrum. They showed that the spectrum is sensitive to even small perturbations in the timing

and amplitudes of glottal pulses. In addition, they illustrated important characteristics that can arise in the spectrum, including apparent shifting of the harmonics and the appearance of multiple pitches. Lee *et al.* (2008) proposed Higher-Order Statistics (HOS) based features to improve classification performance of voice quality measurement. These features were means and variance of kurtosis and skewness which showed meaningful differences in normal, breathy and rough voices. They also used conventional features like jitter, harmonic to noise ratio. They used Classification and Regression Tree (CART) analysis and by utilizing both conventional and HOS based features they were able to get an 89.7% classification rate.

The first voice quality dependent speech recognizer was developed by Yoon *et al.* (2008). They used both spectral and temporal features specifically the harmonic structure and the mean autocorrelation ratio. These features were used by SVM (Support Vector Machine) to classify different type of voice qualities and it was found that Voice quality distinction reflected in PLP coefficients. They obtained 69.23% classification accuracy where the baseline accuracy was 50% a 19% improvement was found which in turn suggests that they could conduct a speech recognition experiment that utilizes the voice quality information, using PLP coefficients as input feature vectors. Finally they used HMM for recognition where first they developed a baseline system and tested its accuracy which acted as the baseline accuracy and then they incorporated the voice quality features into the HMM and tested it. Finally they found that in there was slight increase in the accuracy of the detection.

Based on the literature survey done in this work, it is concluded that very less work has been done for automation of identification of different voice qualities. The cases in which this has been carried out, the algorithms that have been developed identify only a specific type of voice quality. So, in this thesis, an attempt has been made to propose an algorithm that automates simultaneous identification of different voice qualities, *i.e.*, modal, creaky and breathy with a reasonable accuracy.

CHAPTER 2

DATA COLLECTION, PREPROCESSING AND COMPUTATION OF FEATURES

Before detecting the voice quality some work need to be done. First and the most important of this is data collection. After successful collection of data some preprocessing may be required to be done before features are extracted from the data and finally features are extracted on which some algorithms or computation techniques are applied to detect different types of voice quality.

This chapter deals with the three phases of voice quality detection, namely, data collection, preprocessing and feature extraction. Section 2.1 concentrates on Data collection, Section 2.2 includes the preprocessing techniques applied on the sound file and finally section 2.3 discusses the various features used and how they are computed.

2.1 Data Collection

Data in the domain of speech recognition is nothing but utterances of words spoken by speakers. These utterances somehow need to be stored in the digital form. This phase describes how the data is stored in digital form so that further processing can be done on it.

Sound is a waveform having a frequency traveling in any medium. Sound is analog in nature so we need to convert this analog data into digital form so that it can be stored. This is done by a microphone which captures sound waves, by sensing the deflection caused by the wave on a thin membrane, transforming it proportionally to either voltage or current. The resulting electrical signal is converted to a sequence of coded digital data using an Analogue-to-Digital Converter (ADC). Sound is produced if this same sequence of coded data is fed through a compatible digital-to-analogue converter (DAC), through an amplifier to a loudspeaker. The voltage applied to the loudspeaker at instant of time is proportional to the sample value from the computer being fed through the DAC. The voltage on the loudspeaker causes a cone to deflect in or out, and it is this cone which compresses (or rarifies) the air from instant to instant thus initiating a sound

wave. The basic block diagram of digital audio system is shown in the Figure 2.1 (Mcloughlin, 2009).

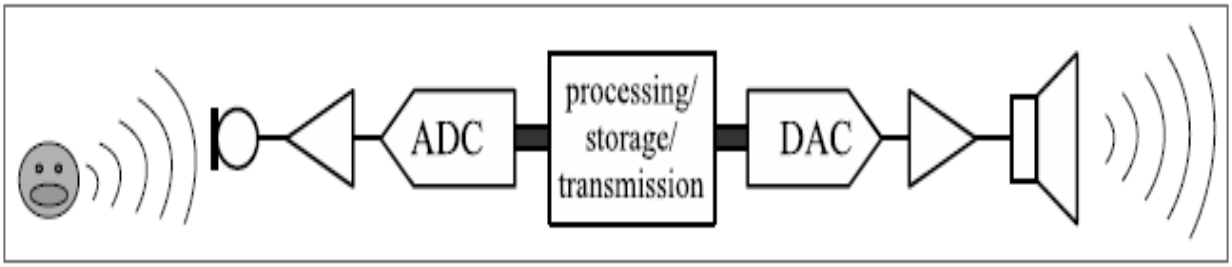


Figure 2.1: Block diagram of a digital audio system.

Now the storage of this digital information can be done in many formats like fixed–points numbers, double precision floating point number etc, but an audio is best handled when stored as a vector with each individual value being a double precision floating point number. Another information is also needed along with the vector of values, *i.e.*, sampling rate or frequency. Any operation that can be performed on a vector can, in theory, be performed on stored audio. The audio vector can be loaded and saved in the same way as any other variable, processed, added, plotted, and so on.

The data that is used is recorded with a sampling frequency of 22050 using the record feature of COLEA tool and 135 words spoken in Hindi by two male speakers have been taken. These words consist of 15 words spoken 3 times by one speaker and 6 times by the other speaker.

The size of sound vector obtained depends on the length of the sound file and the sampling frequency at which the sound has been recorded. If we plot the vector we get an sample vs. amplitude graph. We can also plot a time vs. amplitude graph. These two representations are called Time domain representations. There is another representation called the spectral representation which is an frequency vs. time plot. All the three representations are shown in the Figures 2.2-2.4.

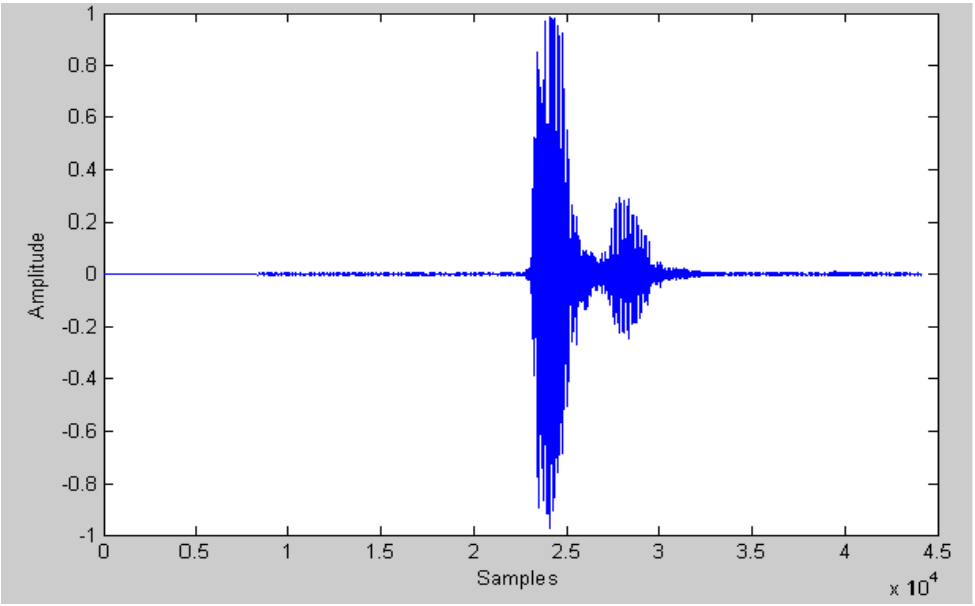


Figure 2.2: A amplitude vs. samples plot of "bahr" in Hindi

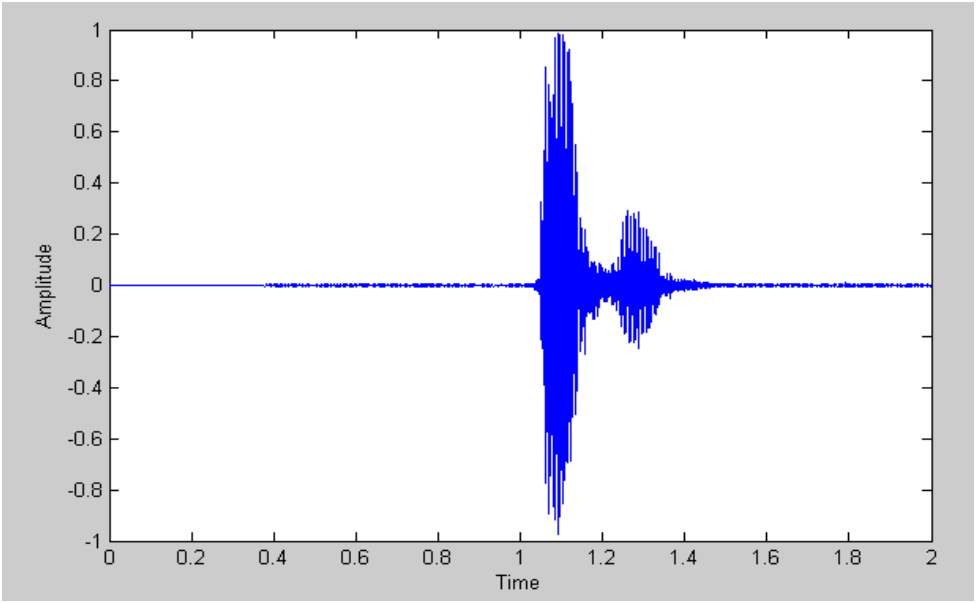


Figure 2.3: A amplitude vs. time plot of "bahr" in Hindi

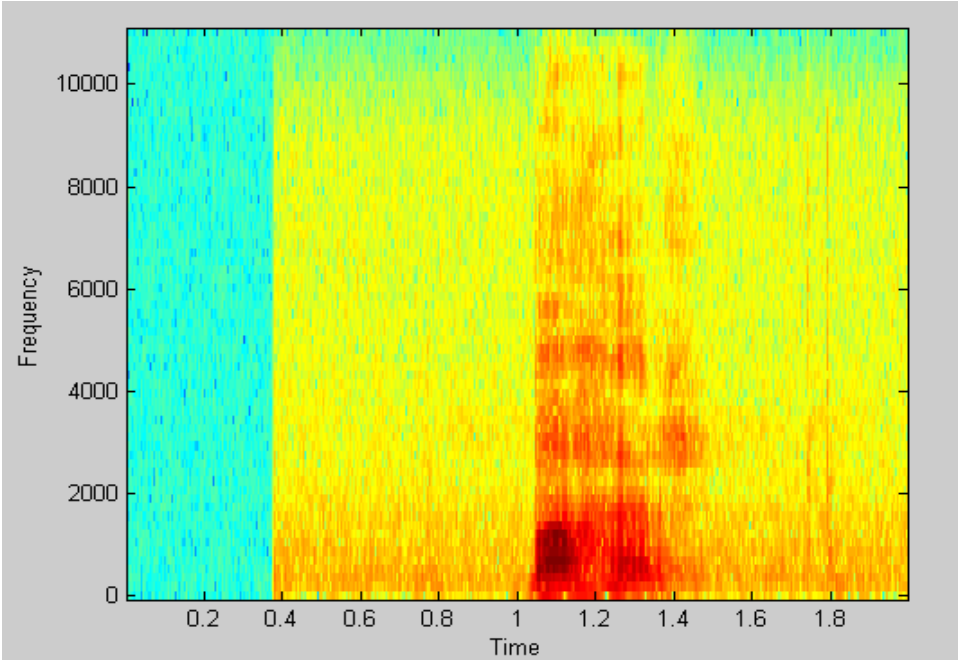


Figure 2.4: Spectrogram (frequency vs. time plot) for "bahr" in Hindi.

2.2 Preprocessing

Data preprocessing is any type of processing performed on raw data to prepare it for another processing procedure. It transforms the data into a format that will be more easily and effectively processed for the purpose of the user. Preprocessing also reduces the amount of work to be done in the further stages. Preprocessing in the domain of speech recognition and its other applications like voice quality detection *etc.* depends upon the work that need to be further done for example a person working on recognizing phonemes from spoken words may see preprocessing as separation of different phonemes in the word, similarly in the case of voiced/unvoiced detection preprocessing may be windowing of the signal *etc.* For the voice quality detection only windowing and framing were used which is explained in the next section.

2.2.1 Windowing and framing

Speech Signal is very random in nature, also the speech vector formed when speech is stored in digitized form is very large in size and depends upon the sampling rate at which the sound is being recorded. A simple formula to find out the length of vector formed is

$$Size = sampling\ freq * duration\ of\ speech\ in\ seconds$$

If the duration of speech is 10 seconds and the sound is recorded at 22050 sampling rate then the size of vector formed will be 2,20,500 and each value will be a double precision floating point number. So to process a vector of such size in one go is very difficult and also speech signal is highly random in nature, so to extract feature is very difficult. To overcome these drawbacks windowing of signal is done, and also for speech processing we want to assume the signal is short-time stationary and perform a Fourier transform on these small blocks. So we multiply the signal by a window function that is zero outside some defined range.

Windowing determines the portion of the speech signal that is to be analyzed by zeroing out the signal outside the region of interest. Windowing techniques include the Rectangular, Bartlett, Hamming, Hanning, Blackman, and Kaiser. The most commonly used are the Rectangular and the Hamming methods.

The rectangular window (sometimes known as the boxcar or Dirichlet window) is the simplest window, equivalent to replacing all but N values of a data sequence by zeros, making it appear as though the waveform suddenly turns on and off.

$$W(n) = 1;\ 0 <= n <= N - 1$$

$$W(n) = 0;\ otherwise$$

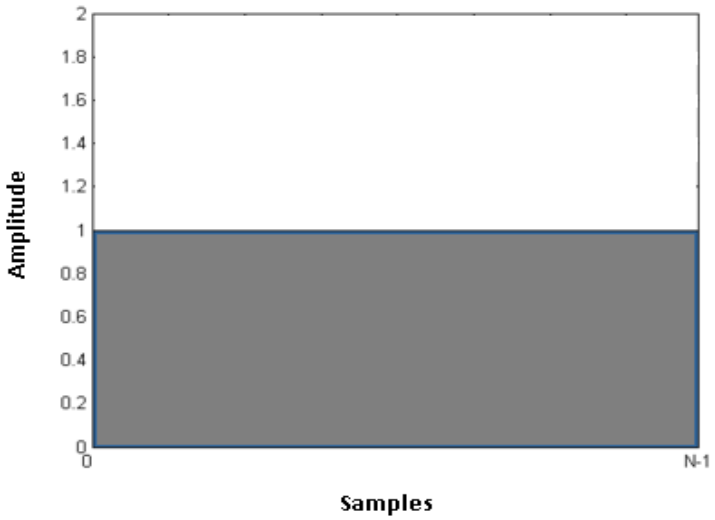


Figure 2.5: Window Function (Rectangular)

Hamming window

This window was proposed by Richard W. Hamming . The window is optimized to minimize the maximum (nearest) side lobe, giving it a height of about one-fifth that of the Hann window, a raised cosine with simpler coefficients.

$$W(n) = 0.54 - 0.46\, cos(2\pi n/N - 1)$$

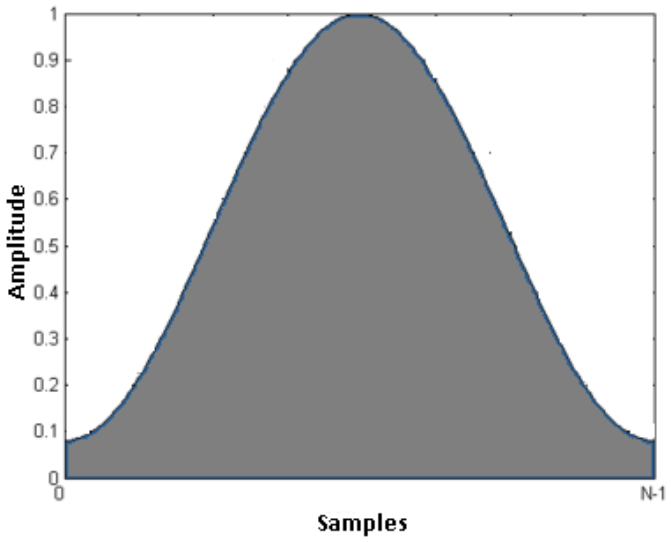


Figure 2.6: Window function (Hamming)

2.3 Feature extraction

Feature extraction involves analysis of speech signal and extraction of information from the speech vector which can be used to differentiate between different types of voice quality. Humans have the ability to easily differentiate between the voice quality of different speakers but to teach this to a machine is a very difficult task. Extracting features which show similar properties for every speaker and conditions is very important task.

Broadly the feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis the speech waveform itself is used for analysis. In spectral analysis spectral representation of speech signal is used for analysis.

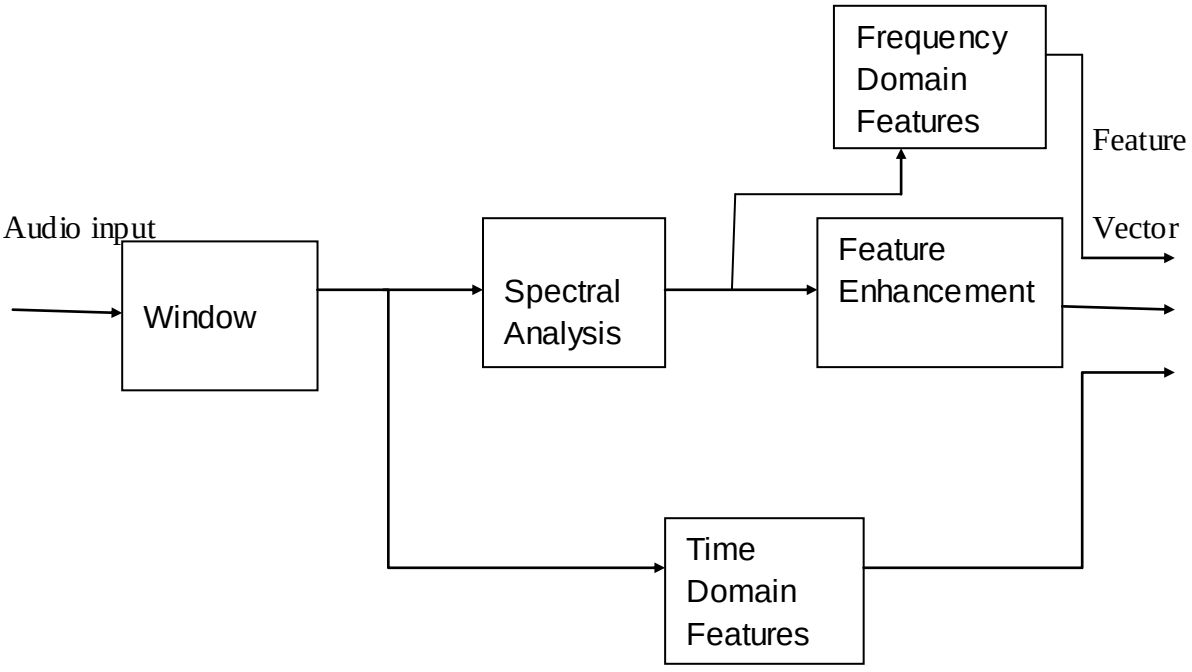


Figure 2.7: Block diagram of feature extractor

Spectral analysis gives us quite a lot of information about the spoken phone. Time domain data is converted to Frequency domain by applying Fourier transform on it. This process gives us the spectral information. Spectral information is the energy levels at different frequencies in a given window. Thus features like frequency with maximum energy, distance between frequencies of maximum and minimum energies *etc*. can be extracted.

Temporal features are easy to extract, simple and have easy physical interpretation. Temporal features like average energy level, zero-crossing rate, root mean square, maximum amplitude *etc*. can be extracted out as features.

For voice quality detection temporal features Zero crossing rate and short time energy along with spectral feature F0 using cepstrum are used.

2.3.1 Fundamental Frequency

Fundamental Frequency (F0) or pitch is defined as the frequency at which the vocal cords vibrate during a voiced sound. Fundamental frequency has long been difficult parameter to reliably estimate from the speech signal. Basically, there are two categories of approaches for pitch tracking (Fundamental Frequency Estimation). One category is in the time domain, and the other category is in the frequency domain. Time-domain analysis could use some time-related features such as ZCR (Zero-Crossing Rate), peak picking, and autocorrelation. Frequency domain analysis could apply, for example, to cepstrum and harmonic matching. These two kinds of approaches have their advantages and disadvantages, respectively; for example, frequency-domain approaches generally have higher accuracy than time-domain methods, but they need more computation. The general method of fundamental frequency estimation is to take a portion of the signal and to find the dominant frequency of repetition (Zaho *et al.*, 2007).

For this work two approaches were tested Time domain approach using autocorrelation and Spectral domain approach using cepstrum with hamming window with 40ms

2.3.1.1 Autocorrelation F0 Detection Approach

Autocorrelation preserves information about harmonic and formant amplitudes in speech signals, while ignoring phase; we use autocorrelation because phase is less important perceptually and

carries much less communication information than spectral magnitude. Actually, our ears are not very sensitive for speech phases. The autocorrelation function is a special case of the cross-correlation function. Assume a speech signal is $s(n)$; its autocorrelation Rss is:

$$R_{ss}(k) = \sum_{m=-\infty}^{\infty} s(m)s(m-k)$$

The autocorrelation measures the similarity of the signal and its time delay. By summing the products of a speech signal and a delayed signal of itself, the autocorrelation is large if at some delay the two signals have similar waveforms. The range of summation is usually limited, and dividing by the number of summed samples could normalize the function. An autocorrelation pitch detector calculates the cross correlation for each block signal with its time delayed signal. In fact, each speech block is measured for similarity with its time delayed signal. If a section of a speech signal is periodic, its autocorrelation function will reach the maximum value at the location of pitch periods.

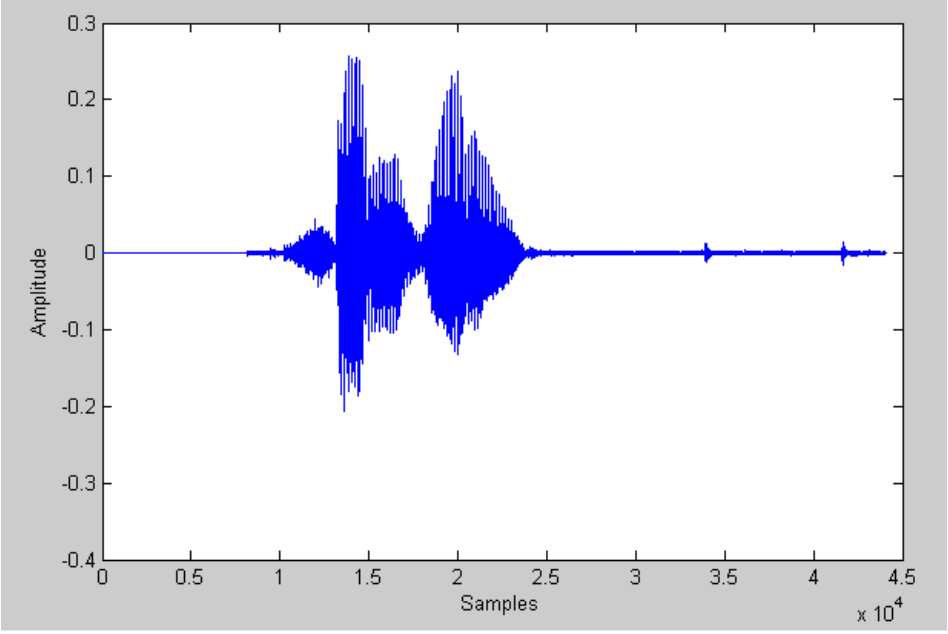


Figure 2.8: Plot of "shalgam"

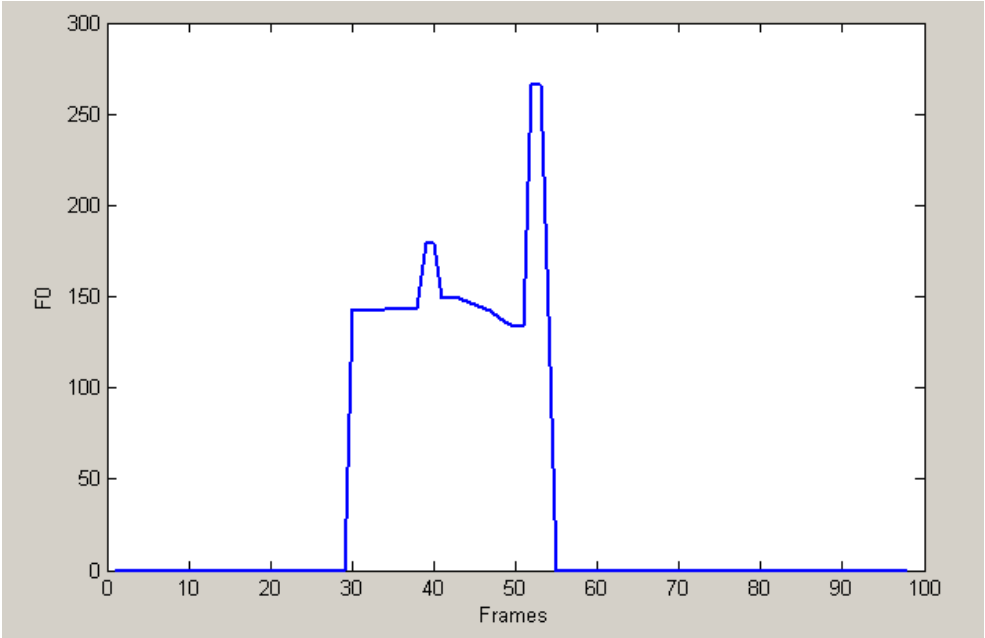


Figure 2.9: F0 of "shalgam" using autocorrelation

2.3.1.2 Cepstral F0 Detection Approach

A reliable way of obtaining an estimate of the dominant fundamental frequency for long, clean, stationary speech signals is to use the cepstrum. 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. The cepstrum is so-called because it turns the spectrum inside out. The cepstrum has units of frequency, and peaks in the cepstrum (which relate to periodicities in the spectrum) are called harmonics. To render the cepstrum suitable for digital algorithms, the DFT must be used in place of the general Fourier transform in Equation:

$$C_d(n) = \frac{1}{N}\sum_{K=0}^{N-1} \log|X(k)|\ e^{j2\pi kn/N}$$

To obtain an estimate of the fundamental frequency from the cepstrum we look for a peak in the frequency region corresponding to typical speech fundamental frequencies.

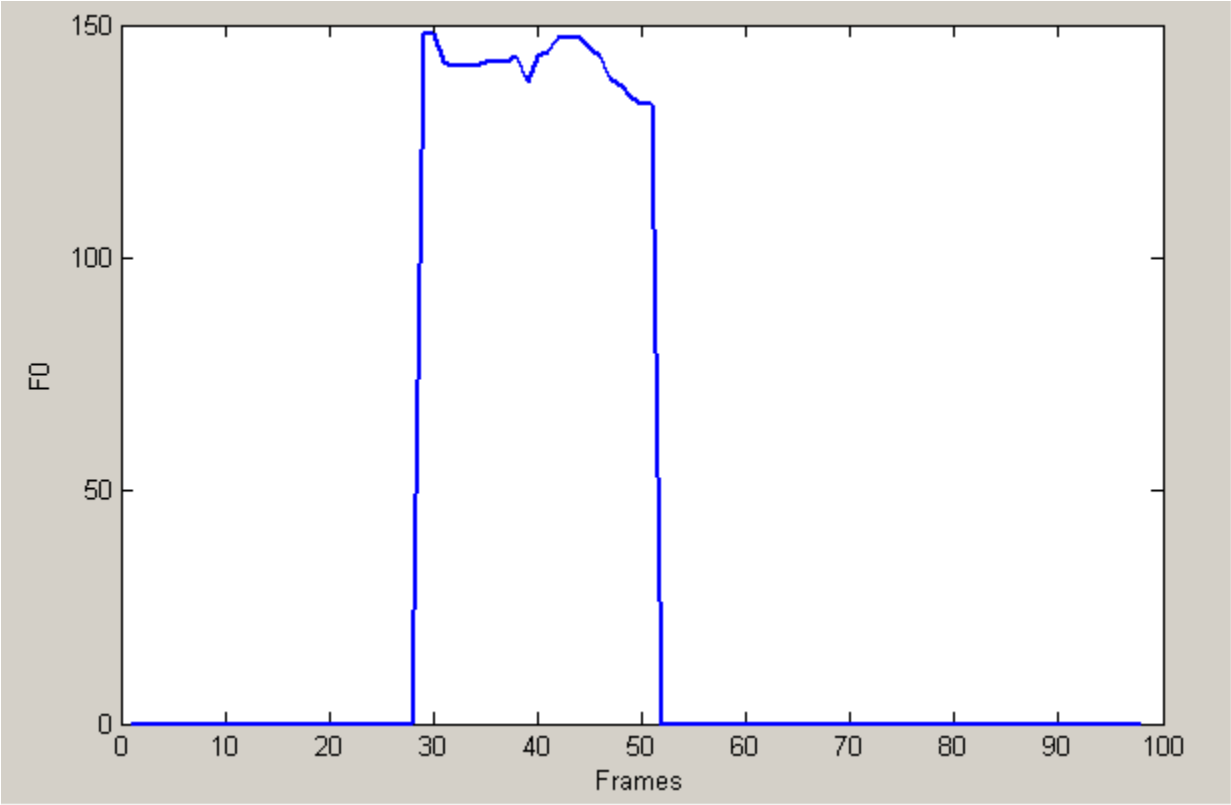


Figure 2.10: F0 "shalgam" using cepstrum

2.3.2 Zero Crossing rate

In mathematical terms, a "zero-crossing" is a point where the sign of a function changes (*e.g.*, from positive to negative), represented by a crossing of the axis (zero value) in the graph of the function.

Counting zero-crossings is also a method used in speech processing to estimate the fundamental frequency of speech. But it is used only for monophonic tone signals not for speech signal where the signal is highly random in nature.

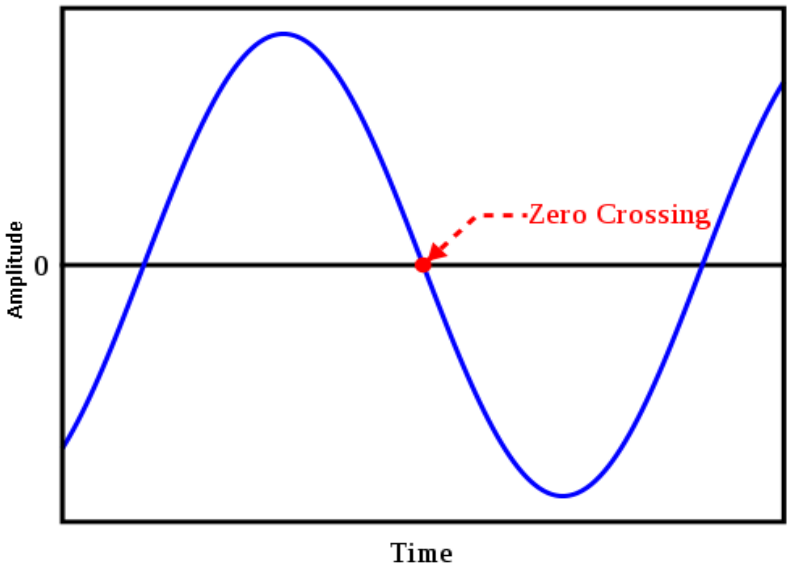


Figure 2.11: Zero Crossing in a waveform

The Zero Crossing Rate is the rate of sign-changes along a signal, *i.e.*, the rate at which the signal changes from positive to negative or back (Rabiner and Schafer, 2007). This feature has been used heavily in both speech recognition and music information retrieval.

ZCR is defined formally as

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

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 Zn is a weighted sum of all the instances of alternating sign (zero-crossing) that fall within the support region of the shifted window $w[\hat{n} - m]$.

For this work to calculate Zero Crossing Rate the signal is first windowed using rectangular window with 20 ms length and then zero crossing rate is calculated.

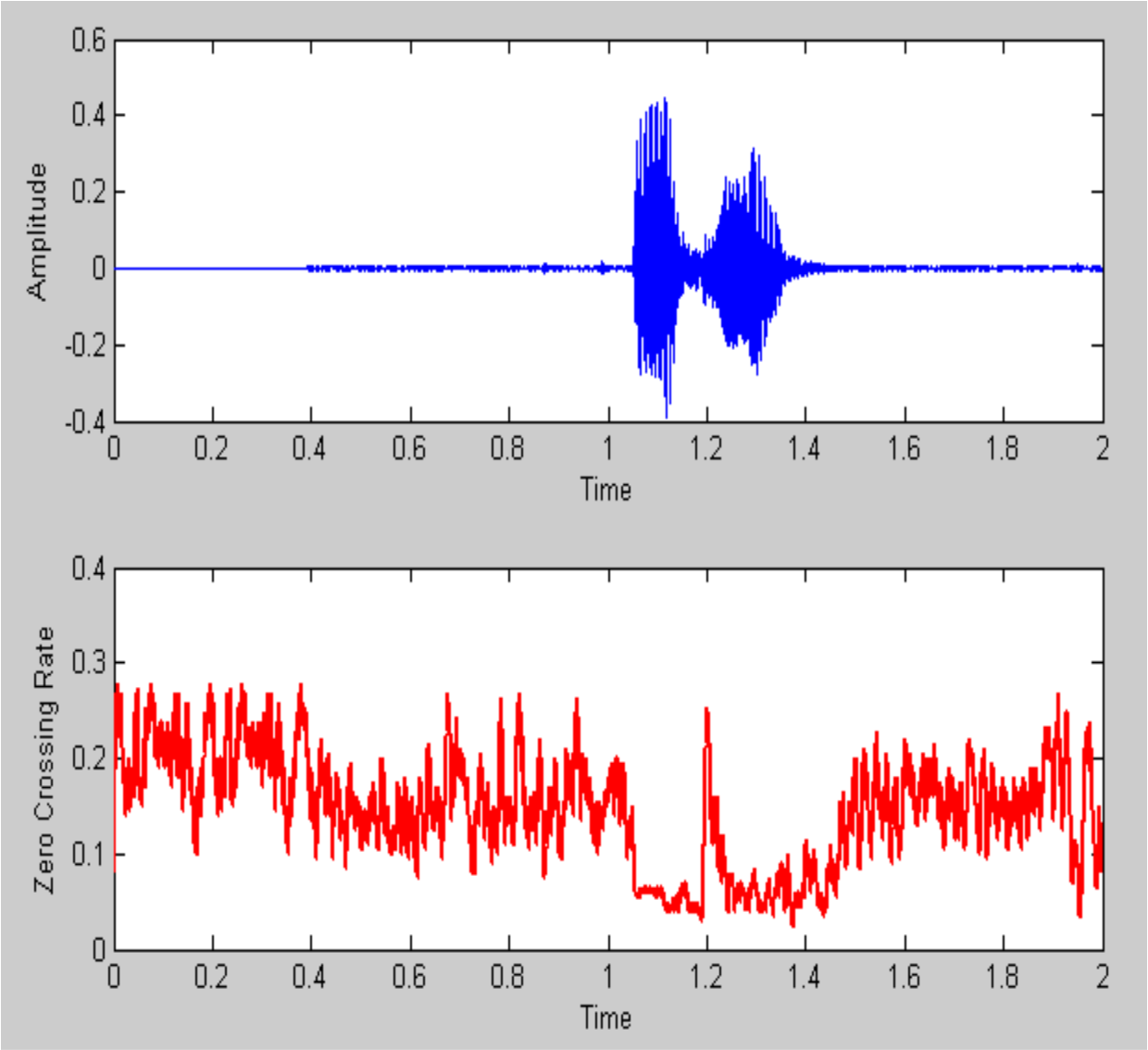


Figure 2.12: Speech waveform for word "ghar" and its Zero Crossing Rate

2.3.3 Short Time Energy

Short-Time Energy (STE) is the energy associated with the signal in time domain. In order to extract the short time energy from the signal the signal was first broken into non-overlapping

short-term-windows (frames) of 50 ms each. Then for each frame the short time energy was calculated.

Energy is calculated using the following formulae

$$E(i) = \frac{1}{N} \sum_{n=1}^{N} |x_i(n)|^2$$

Where $xi(n);\ n = 1..N$ is the audio samples of the i[th] frame, of length N.

For windowed sample this formula changes to

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

where, En = Short-Time Energy, x = Speech Signal, w =Window.

After the calculation of short time energy threshold is calculated. Threshold can be either decided statically which is normally taken as 30 db but dynamic calculation of threshold is more effective (http://www.mathworks.com/matlabcentral/fileexchange/28826-silence-removal-inspeech-signals). To calculate the threshold following steps are followed:-

Step 1:- Compute the histogram of the feature sequence's values.

Step 2:- Detect the histogram's local maxima.

Step 3:- Let *M*1 and *M*2 be the positions of the first and second local maxima respectively. The threshold value is computed using the following equation:

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

Here, *W* is a user-defined parameter. Large values of *W* obviously lead to threshold values closer to *M*1.

Step 4:- If only one local maxima is found then *T* is taken simply as half of the mean of energy.

This threshold obtained successfully shows the difference where the there is spoken word and where not.

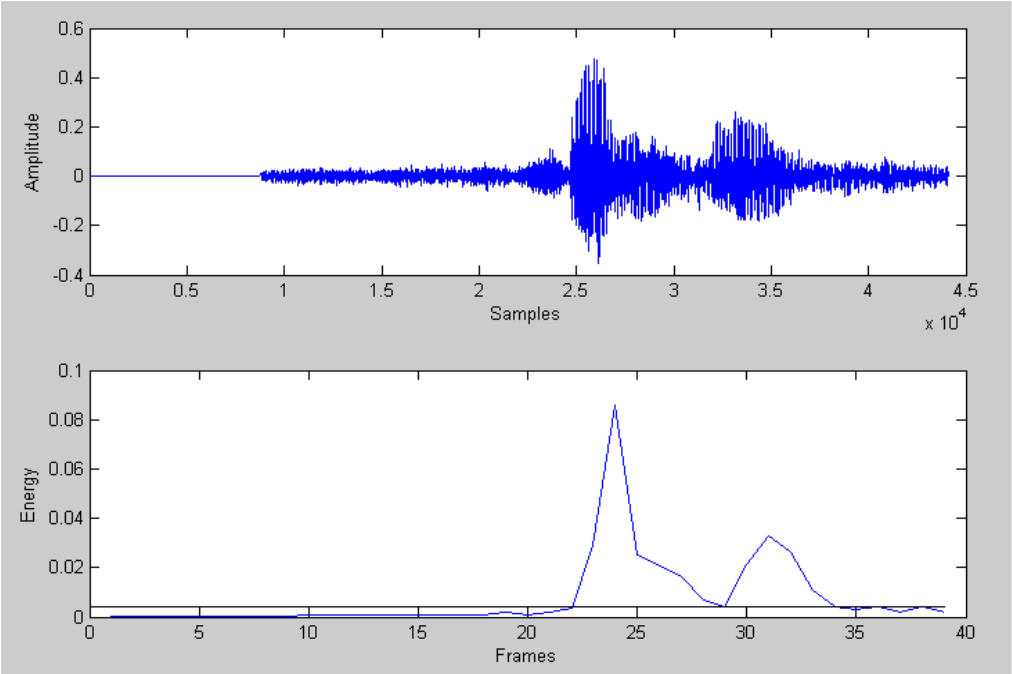


Figure 2.13: Speech Waveform and its short time energy with thresholding.

For calculating the short time energy 50 ms window was best suited because if we chose 20 ms second the energy plot obtained was not very smooth and showed lot of variation.

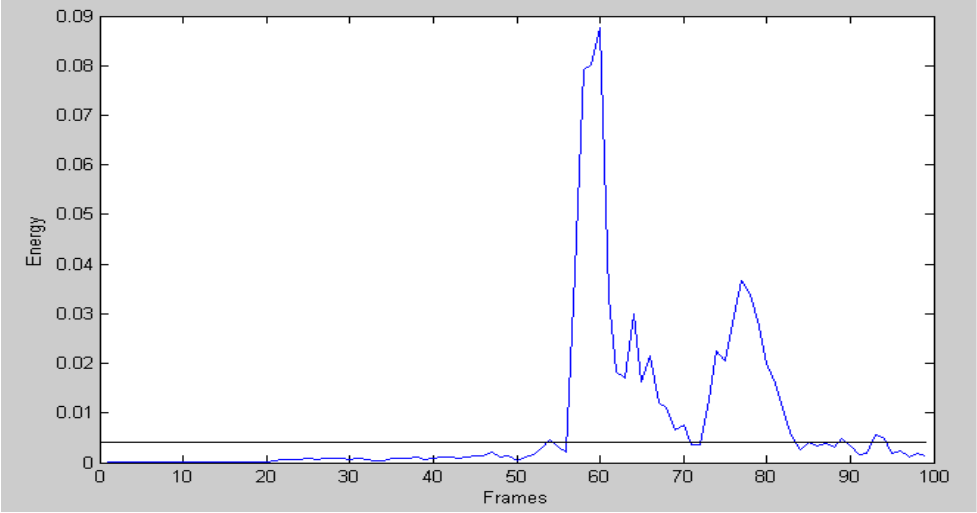


Figure 2.14: Short time energy plot with 20 ms window

On the other hand if we take 100 ms window the energy plot obtained showed very less variation and also some pauses in the speech signal were not detected.

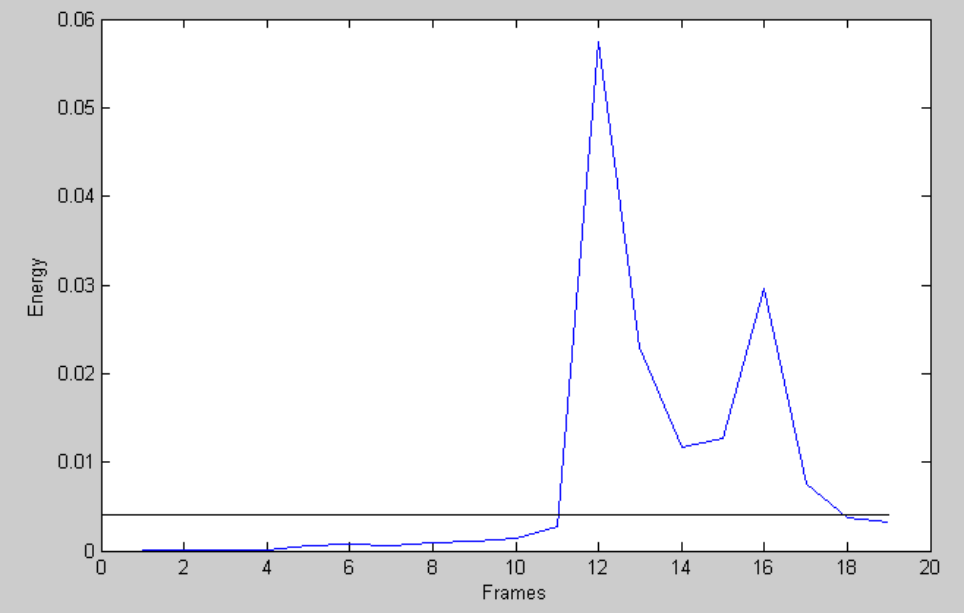


Figure 2.15: Short time energy plot with 100 ms window

All the above three features ,*i.e.*, F0, Zero Crossing rate and Short time energy were calculated for the speech signal and their feature vector were maintained.

The next chapter describes how these features are used and presents the algorithm proposed to identify modal, creaky and breathy voices.

CHAPTER 3

CLASSIFICATION OF CREAKY, MODAL AND BREATHY VOICE

Speech or sound is produced when the lungs expel sufficient amount of air through glottis which create a pressure drop across the larynx. Due to this pressure the vocals folds start to oscillate which produces sound. The voice quality of a person depends upon the state of glottis when the person is speaking (Titze, 1994). Basically there are three types of voice qualities found: modal, creaky, breathy.

Modal voice quality is produced when the length, tension, and mass of the vocal folds are in a state of flux which causes the frequency of vibration of the vocal folds to vary. Breathy voice quality is produced when the vocal cords vibrate, as they do in normal (modal) voicing, but are held further apart, so that a larger volume of air escapes between them. This produces an audible sound. Creaky voice (pulse phonation, vocal fry, or glottal fry) quality is produced when the vocal fords are compressed tightly, becoming relatively slack and compact. Normally creaky voices occur at low pitched but they can also occur at high pitch (Yoon, 2009). But these voice qualities do not affect the unvoiced part of sound and it is always considered to be modal.

In many languages like Mazatec, Javanese *etc*. the same word when spoken in different voice quality can have different meaning (Gordan, 2001). Also in other languages voice quality factor is very important factor. Human beings can easily identify what is spoken even when the same word is spoken in different voice qualities but it is not possible for the machine to identify the same word when spoken in different voice qualities. To achieve this goal we need to incorporate voice quality detection technique into the speech recognition.

The following sections describe the facts and results obtained from the features used and then present an algorithm for identification of voice quality and finally results are shown and discussed.

3.1 Results and facts obtained from features

To identify different voice qualities three features were used which are explained in the previous chapter. Some facts were observed from the features which are discussed in the next section.

3.1.1 Zero Crossing Rate (ZCR)

It was observed that both modal and creaky voice qualities the ZCR was less than .1 but for breathy voice and noise it was above .1 and less than .3 and for voiceless sounds it was above .3.

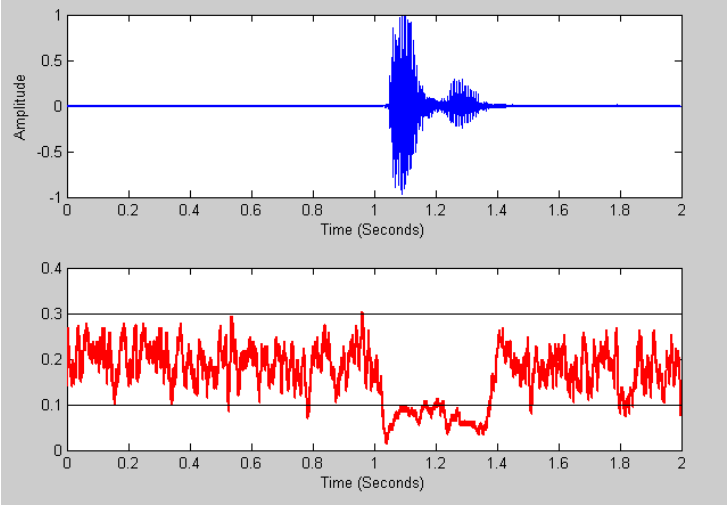


Figure 3.1: Plot of "bahr" spoken in modal voice quality with its zero crossing rate.

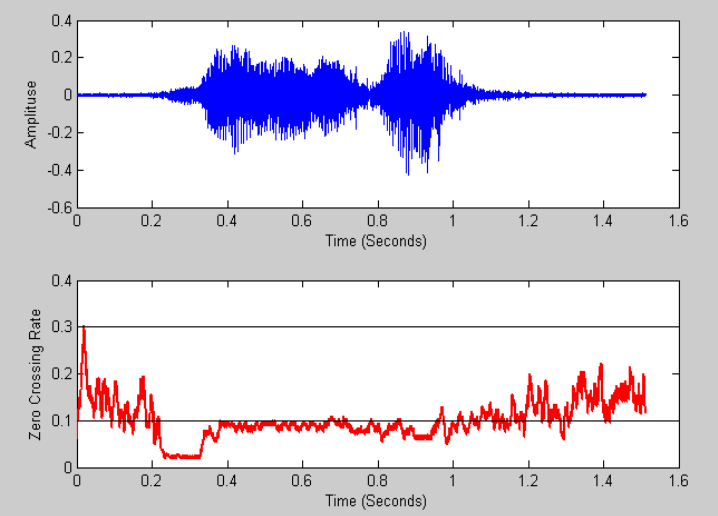


Figure 3.2: Plot of "bahr" spoken in creaky voice quality with its zero crossing rate.

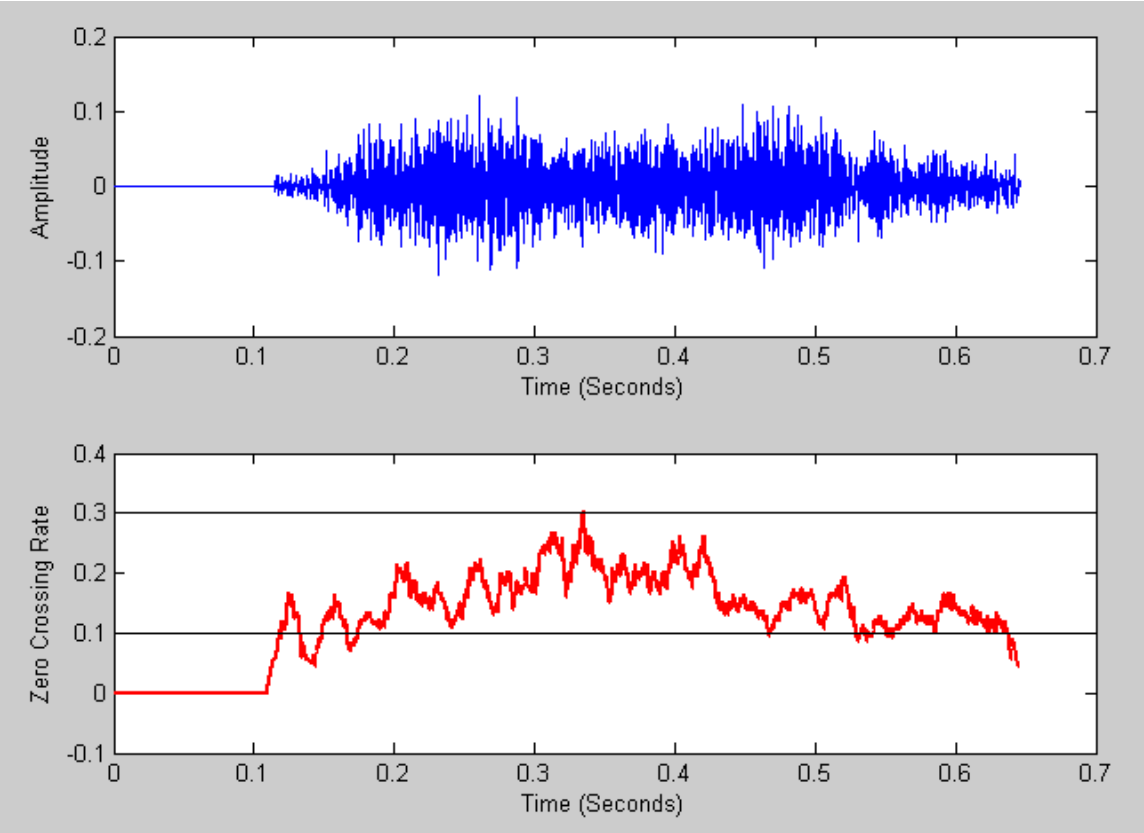


Figure 3.3: Plot of "bahr" spoken in breathy voice quality with its zero crossing rate.

From the above results it can be concluded that zero crossing rate can only be used to divide or classify the signal into two parts ,*i.e.*, modal and creaky or breathy and noise. To further identify we need to incorporate some other feature.

3.1.2 Fundamental Frequency (F0)

To differentiate between the modal voice and creaky sound fundamental frequency estimation (also called pitch detection) is used. It is well known that F0 of creaky sound is always less than modal sound because for modal voice vocal fold length may be considered to be medium with the length increasing as fundamental frequency (F0) increases and for vocal fry (creaky voice)

vocal fold length is short which causes the F0 to decrease from that of modal voice (Childers and Lee, 1991).

Tests were done to find out the range of F0 and it was observed that the F0 of creaky sound always lied in the range of 95 to 135 Hz and for modal voice it was above 135 Hz.

Table 3.1 shows the average F0 value obtained for the subset of creaky data set.

Table 3.1: Average F0 for creaky words

S. No.	Word spoken	Average F0
1	"ajay" (1)	125.4
2	"ajay" (2)	116.9
3	"ajay" (3)	121.8
4	"bahr" (1)	129.2
5	"bahr" (2)	126.4
6	"bahr" (3)	130.6
7	"ghar" (1)	127.6
8	"ghar" (2)	128.5
9	"ghar" (3)	132.4
10	"kabutar" (1)	126.1
11	"kabutar" (2)	130.1
12	"kabutar" (3)	123.4
13	"shalgum" (1)	123.3
14	"shalgum" (2)	133.4
15	"shalgum" (3)	130.4
16	"aag" (1)	114.0
17	"aag" (2)	117.8

18	“aag” (3)	118.7
19	“chabhi” (1)	103.7
20	“chabhi” (2)	102.8
21	“chabhi” (3)	100.0
22	“chori” (1)	112.4
23	“chori” (2)	119.2
24	“chori” (3)	121.6
25	“gamla” (1)	113.6
26	“gamla” (2)	107.1
27	“gamla” (3)	113.0
28	“ghas” (1)	113.0
29	“ghas” (2)	107.4
30	“ghas” (3)	109.0
31	“kagaj” (1)	120.2
32	“kagaj” (2)	118.2
33	“kagaj” (3)	124.0
34	“kalam” (1)	120.6
35	“kalam” (2)	117.9
36	“kalam” (3)	114.4
37	“ped” (1)	113.0
38	“ped” (2)	111.0
39	“ped” (3)	115.5
40	“shor” (1)	108.2
41	“shor” (2)	119.9
42	“shor” (3)	121.3

43	“suraj” (1)	122.7
44	“suraj” (2)	122.0
45	“suraj” (3)	124.2

It is clear from the above data that average F0 lies in the range of 95 to 135 Hz. There may be some exceptions but for this data set the range was correct.

Table 3.2 shows the average F0 value of the subset of modal data set

Table 3.2: Average F0 for modal words

S. No.	Word spoken	Average F0
1	“ajay” (1)	153.0
2	“ajay” (2)	153.9
3	“ajay” (3)	155.4
4	“bahr” (1)	163.1
5	“bahr” (2)	141.1
6	“bahr” (3)	150.5
7	“ghar” (1)	155.8
8	“ghar” (2)	147.8
9	“ghar” (3)	146.9
10	“kabutar” (1)	144.5
11	“kabutar” (2)	157.3
12	“kabutar” (3)	166.6
13	“shalgum” (1)	145.8
14	“shalgum” (2)	141.8
15	“shalgum” (3)	155.7
16	“aag” (1)	160.2
17	“aag” (2)	161.1

18	“aag” (3)	163.5
19	“chabhi” (1)	180.9
20	“chabhi” (2)	185.6
21	“chabhi” (3)	182.9
22	“chori” (1)	190.1
23	“chori” (2)	157.5
24	“chori” (3)	188.4
25	“gamla” (1)	171.9
26	“gamla” (2)	172.9
27	“gamla” (3)	173.1
28	“ghas” (1)	152.6
29	“ghas” (2)	151.4
30	“ghas” (3)	151.0
31	“kagaj” (1)	171.4
32	“kagaj” (2)	170.5
33	“kagaj” (3)	170.3
34	“kalam” (1)	169.3
35	“kalam” (2)	166.8
36	“kalam” (3)	170.2
37	“ped” (1)	149.4
38	“ped” (2)	151.2
39	“ped” (3)	155.5
40	“shor” (1)	167.4
41	“shor” (2)	161.8
42	“shor” (3)	164.6
43	“suraj” (1)	180.2
44	“suraj” (2)	186.1
45	“suraj” (3)	187.6

For modal voice the average F0 obtained is always greater than 140 Hz. For words containing unvoiced parts it was a near 140 and in some cases it may be a bit less than 140, so range 135 Hz was decided as the lower limit for modal voices.

The following figures show the difference in F0 obtained when a word "bahr" was spoken in two different voice qualities, modal and creaky. The black line in the frames vs. F0 graph shows the maximum range for creaky voice (135 Hz).

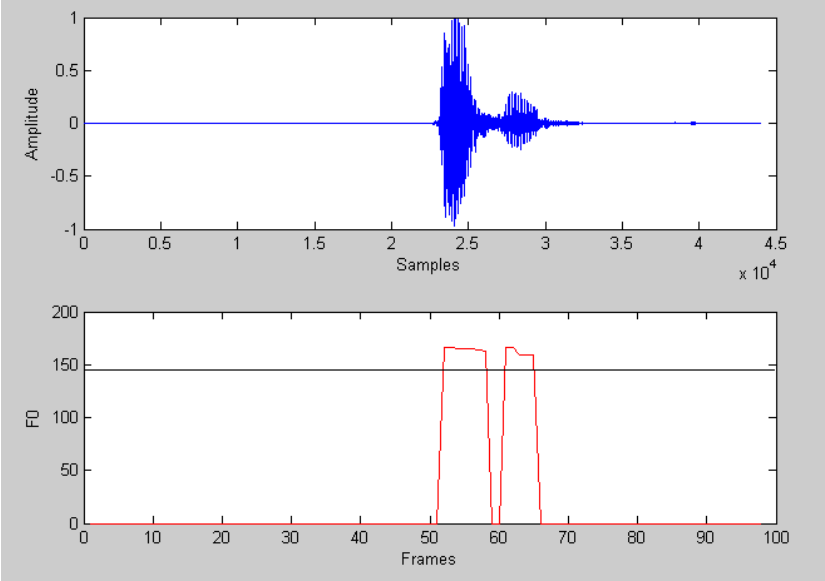


Figure 3.4: Plot of "bahr" spoken in modal voice along with its F0

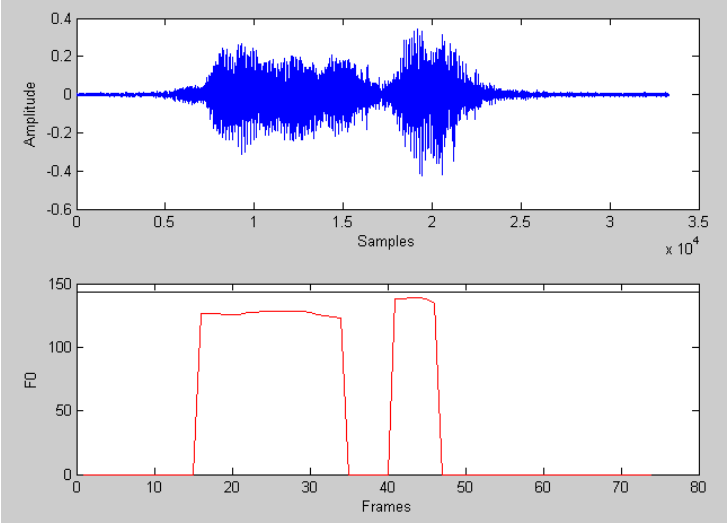


Figure 3.5: Plot of "bahr" spoken in creaky voice along with its F0

With the help of F0 modal and creaky sounds are easily differentiated but breathy sound was not separable using this technique. So another feature Short Time Energy is used.

3.1.3 Short Time Energy (STE)

As discussed previously that breathy sound and noise have same zero crossing rate so using zero crossing rate they are inseparable. But the STE of noise is very less than of voiced region and breathy sound lies in the voiced region (Atal and Rabiner, 1976; Gerratt and Kreiman, 2001). This property of breathy sound is used for the identification of breathy sound.

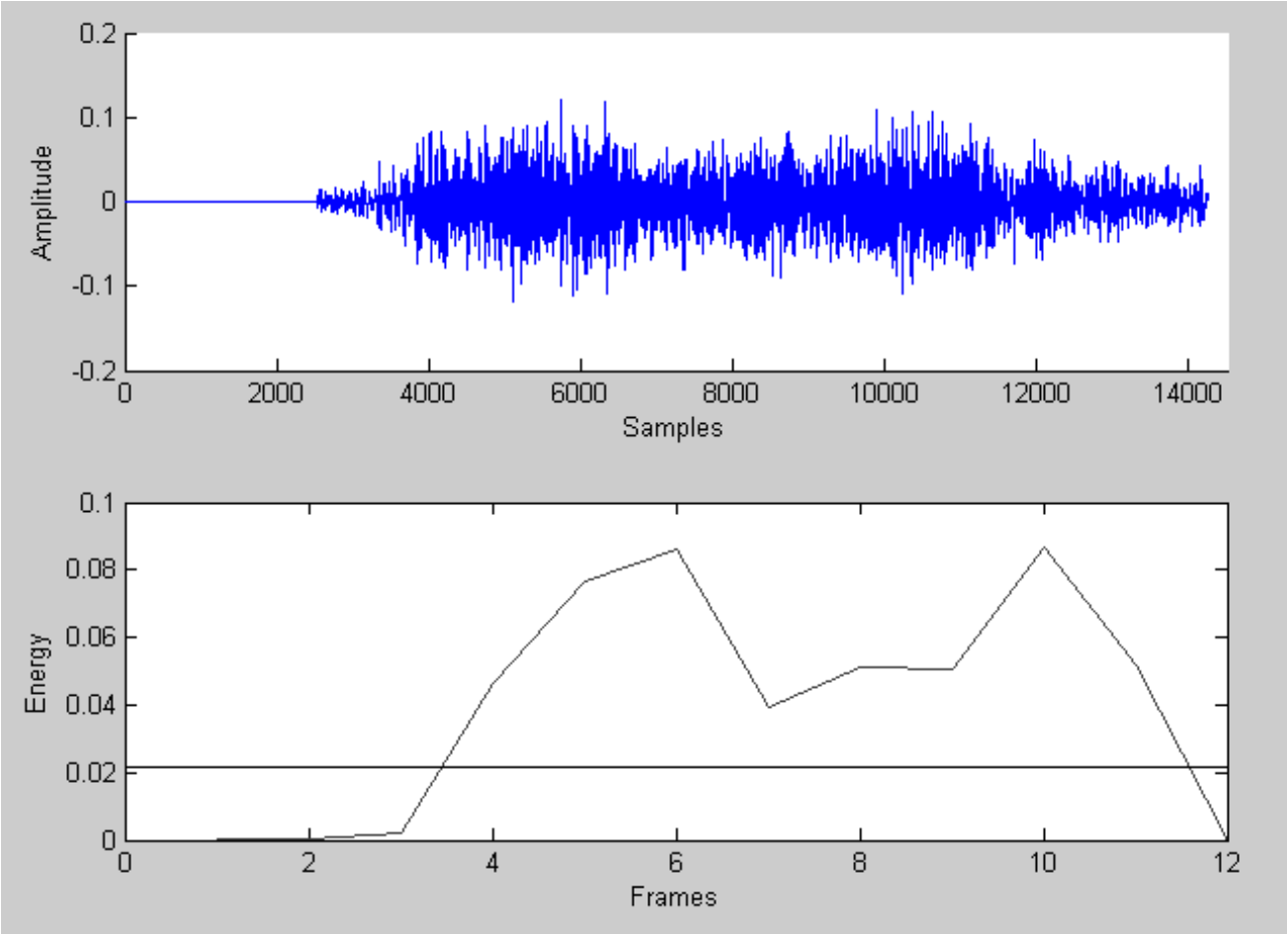


Figure 3.6: Plot of "bahr" spoken in breathy voice along with the plot of energy

3.2 Algorithm used for identification of modal, creaky and breathy voices

Based on the facts and result obtained from the above discussed features an algorithm was designed to identify the modal, creaky and breathy voices. The subsequent sections discuss the algorithm used and show the flowchart.

3.2.1 Algorithm

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

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

Step 3: Compute Fundamental Frequency Vector (F) 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 X and initialize all its values to zero.

Step 9: Repeat for $i = 1\ to\ n$

If $ZC(i) < 0.1$

If $F(i) > 145$ then

set $OUT(i) = 0.1$ (for modal sound)

else

set $OUT(i) = 0.3$ (for creaky sound)

end if

Else if $ZC(i)$ between 0.1 and 0.3 then

If $STE(i) > T\_E$ then

then set $OUT(i) = 0.2$ (for breathy sound)

Else if $ZC(i) > 0.3$ then

set $OUT(i) = 0.1$ (for modal sound)

End if

Step 10: plot X and OUT.

3.2.2 Flowchart

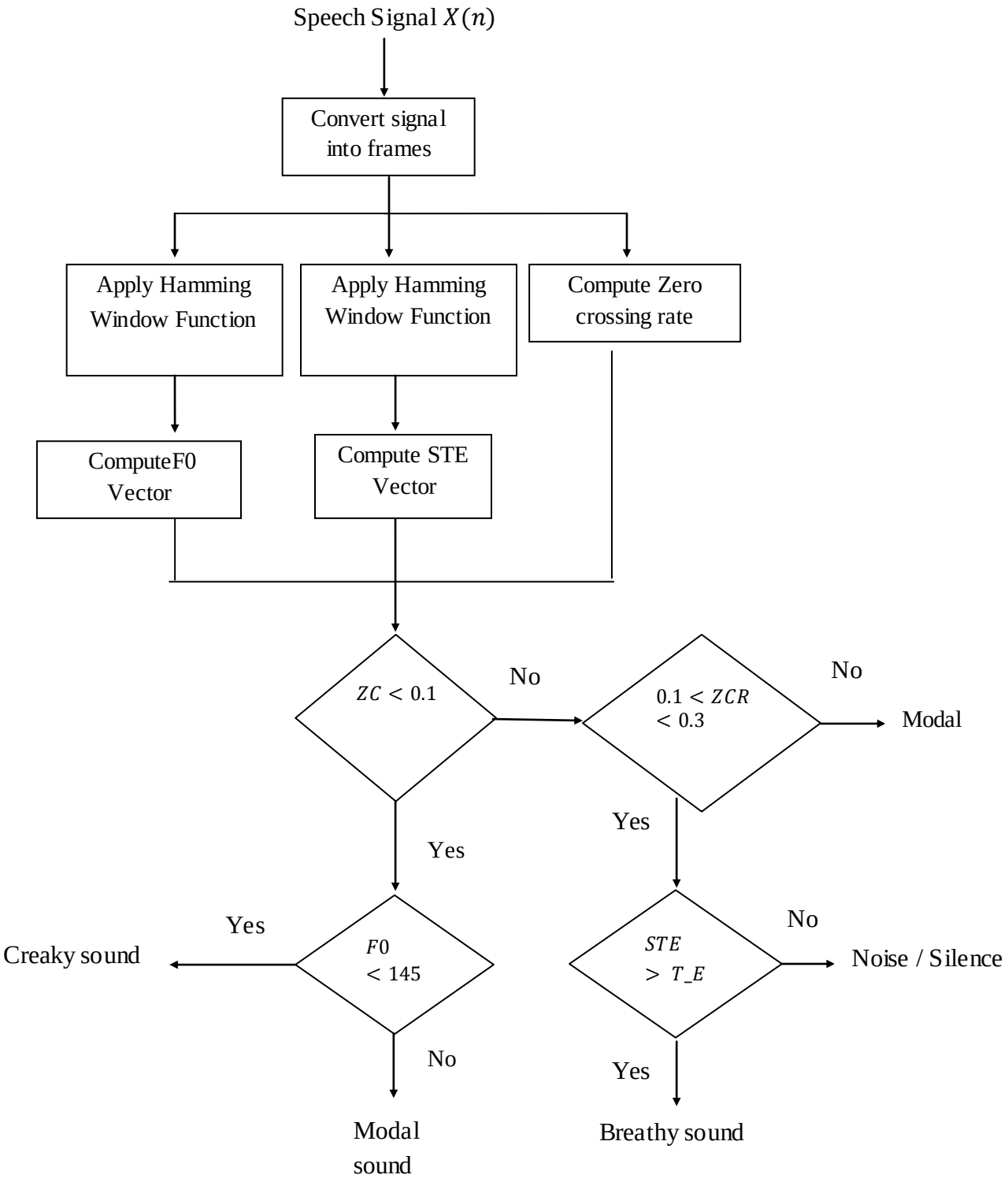


Figure 3.7. : Flowchart of algorithm

The above algorithm was implemented in MATLAB 2011a and then applied to the data. The results were taken in the form of an array and specific values were considered for modal (as 0.1), breathy (as 0.2) and creaky (as 0.3) to be stored in the array. The results and accuracy of algorithm obtained are discussed in detail in the next chapter.

CHAPTER 4

RESULTS AND DISCUSSIONS

This chapter discusses the results obtained after applying the algorithm which is proposed in Chapter 3. Section 4.1 shows the various outputs obtained when the algorithm was applied to different words spoken in different voice qualities. Section 4.2 discusses the accuracy obtained from the algorithm.

4.1 Output of algorithm

The output of algorithm is an out vector which contains different values depending upon the voice quality of the spoken frame. This vector was plotted as a function of the signal samples to get a clear view of where there is modal, creaky or breathy voice.

Figure 4.1 shows a plot of "ghar" spoken in breathy voice along with the zero crossing rate and out vector plotted over the signal.

0.1 value of the output vector means that the region is modal, 0.2 value means that region is breathy, 0.3 value means that it is silence/noise region ,*i.e.,* nothing is being spoken in this region and finally 0.4 value means that it is creaky region.

The signal is plotted using red color, ZCR is plotted using green color and finally the output vector is plotted using blue color.

As it is clear from the figure the word is spoken in breathy voice but a small spike is occurring in plot of output in the breathy region here the value of ZCR is less the 0.1 which is causing the region to be identified as modal region. This region is wrongly detected and is causing a small percentage of error.

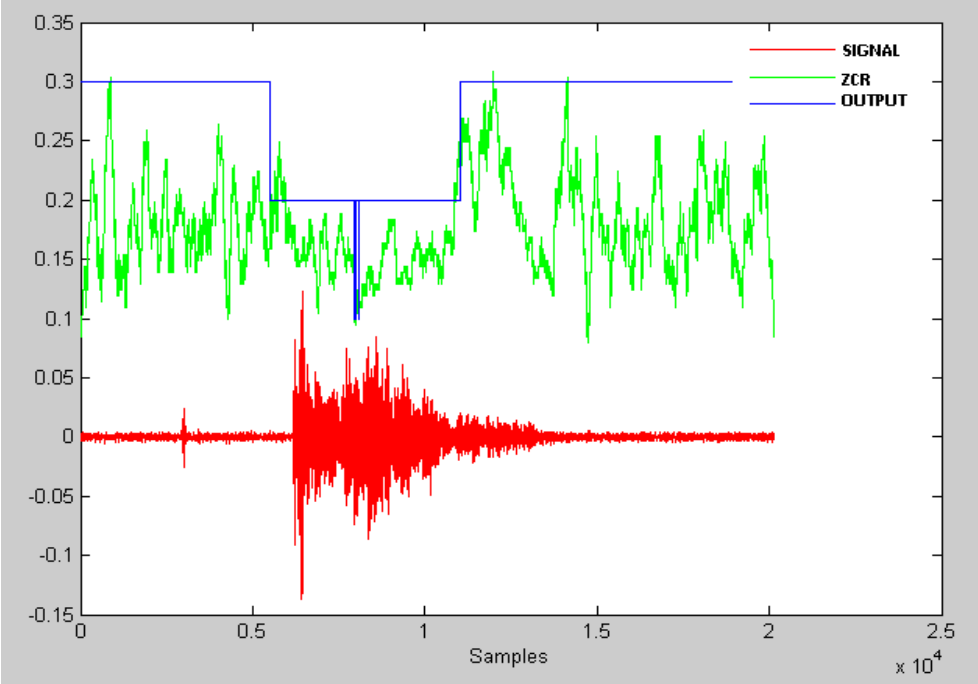


Figure 4.1: Plot of "ghar" spoken in breathy voice along with zero crossing rate and output vector

Figure 4.2 shows the plot of energy of the previous signal, i.e., "ghar" spoken in breathy voice. It can be clearly seen that the energy of breathy region is greater than the silence region.

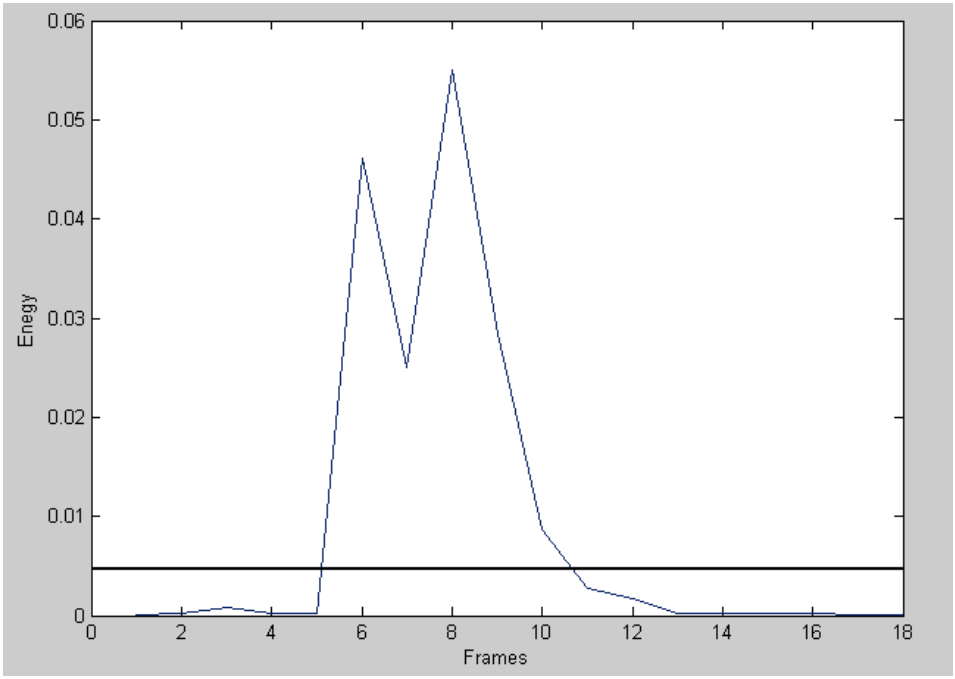


Figure 4.2: Plot of energy for "ghar" spoken in breathy voice

Figure 4.3 shows the output obtained when the algorithm was applied to “ajay” spoken in modal voice. The algorithm detected it as modal with some errors in identification.

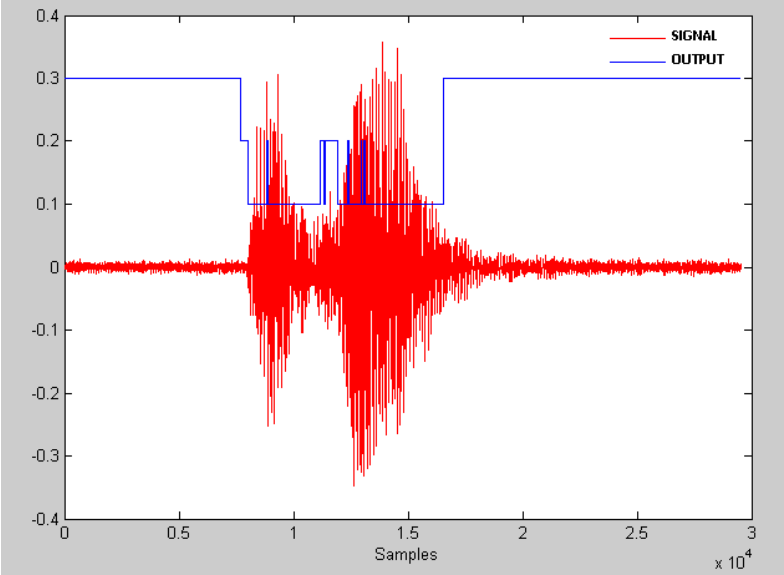


Figure 4.3 Output of algorithm for “ajay” spoken in modal voice.

Figure 4.4 shows the output for the word “ghar” spoken in creaky voice most of the part was correctly identified as creaky but the ending of the word was identified as modal.

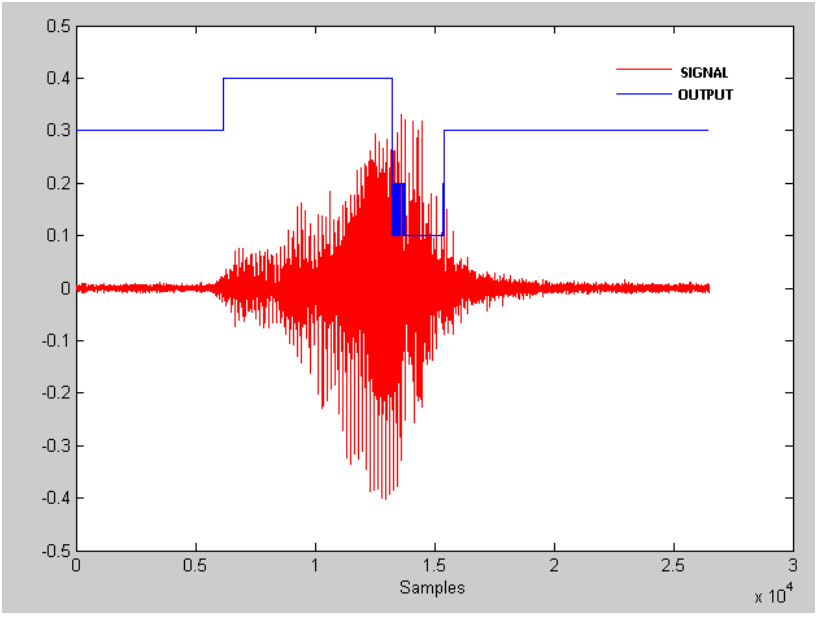


Figure 4.4 Output of algorithm for “ghar” spoken in creaky voice

4.2 Results obtained from algorithm

The previous section showed the output obtained from the algorithm in the form of graphs. This sections deals with the accuracy of the identification of different voices.
The accuracy is calculated as

$$Accuracy\ (in\ percentage) = \frac{No.of\ samples\ correctly\ detected}{Total\ voiced\ samples} * 100$$

Table 4.1 shows the accuracy obtained for 15 words spoken 3 times each in modal voice and the average accuracy obtained.

Table 4.1: Accuracy (in percentage) obtained from algorithm for modal voice.

Sr. No	Word	Accuracy (1st time)	Accuracy (2nd time)	Accuracy (3rd time)	Average accuracy (in percentage) for each word
1	"ajay"	91.7	91.1	90.71	91.17
2	"ghar"	94.4	81.8	90.9	89.03
3	"bahr"	96	89.6	90.89	92.16
4	"kabutar"	93	84.6	85.6	87.73
5	"shalgum"	86.6	81.5	88.6	85.56
6	"aag"	94.1	93.7	97.3	95.03
7	"chabi"	89.3	92.7	86.7	89.56
8	"chori"	93.5	81.2	85.33	86.67
9	"gamla"	95.6	96.7	94.5	95.6
10	"ghas"	94.3	96.3	92.45	94.35

11	“kagaj”	86.3	85.78	84.7	85.59
12	“kalam”	96.4	97.32	94.56	96.09
13	“ped”	92.24	93.45	87.5	91.06
14	“shor”	93.67	94.45	91.24	93.12
15	“suraj”	94.6	89.12	95.43	93.05
Average accuracy (in percentage) for words spoken in modal voice					91.05

As it is clear from the above table high degree of accuracy is achieved in identifying modal voices. Almost 91.05% of the samples are correctly identified as modal sounds. Accuracy up to 96.4% is achieved for some words and for some words it is a bit low ,*i.e.*, 81.5%

The next table ,*i.e.*, Table 4.2 shows the accuracy rate obtained for 15 words spoken 3 times each in creaky voice.

Table 4.2: Accuracy (in percentage) obtained from algorithm for creaky voice.

Sr. No	**Word**	**Accuracy (1st time)**	**Accuracy (2nd time)**	**Accuracy (3rd time)**	**Average accuracy (in percentage) for each word**
1	“ajay”	79.1	72.1	75.8	75.6
2	“ghar”	80	72.2	75	75.73
3	“bahr”	82.3	75.6	83.3	80.4
4	“kabutar”	75	60	72	69
5	“shalgum”	81.2	77.5	75	77.9
6	“aag”	86.9	84.78	83.3	84.93

7	“chabi”	80.63	76	81.5	79.37
8	“chori”	90	90.3	93.4	91.23
9	“gamla”	76	72.7	89.5	79.4
10	“ghas”	81.2	86.3	83.3	83.6
11	“kagaj”	88	92.8	88.4	89.73
12	“kalam”	86.9	77.3	83.3	82.5
13	“ped”	83.2	72	85.7	80.3
14	“shor”	82.3	68.4	70.5	73.3
15	“suraj”	86.9	91.3	85.6	87.93
Average accuracy (in percentage) for words spoken In creaky voice					80.72

The accuracy obtained for creaky voice is less than that for modal voice. This is due to the mixed nature of creaky voice ,*i.e.*, even if the speaker speaks in creaky voice some phonemes like /ey/, /r/, /ta/ *etc.* are not creaky.

For words like “kabutar” where the phoneme /ta/ is extended for a long time the accuracy rate is 67.5%. The average accuracy obtained is 80.72%.

The accuracy can be improved by considering average F0 of whole word (not the frames) but then the algorithm will work for audio file containing only one word. So there is a tradeoff between accuracy and multiple words in a file.

Table 4.3 contains the accuracy obtained for 15 words spoken 3 times each in breathy voice and the average accuracy obtained.

Table 4.3: Accuracy (in percentage) obtained from algorithm for breathy voice.

Sr. No	Word	Accuracy (1st time)	Accuracy (2nd time)	Accuracy (3rd time)	Average accuracy (in percentage) for each word
1	"ajay"	92.2	88.8	87.5	89.5
2	"ghar"	88	95.2	91.68	91.62
3	"bahr"	90.9	85.9	88.78	88.52
4	"kabutar"	89.9	93.3	84.6	89.26
5	"shalgum"	91.1	92.85	90.1	91.35
6	"aag"	98.86	93.53	94.65	95.68
7	"chabi"	87.52	92.34	86.66	88.84
8	"chori"	94.16	86.35	89.78	90.09
9	"gamla"	93.3	97.2	98.3	96.26
10	"ghas"	97.1	93.75	92.85	94.56
11	"kagaj"	93.75	95.68	91.23	93.55
12	"kalam"	66.6	74.3	75.91	72.27
13	"ped"	93.3	74.22	90.13	85.88
14	"shor"	97.5	90.16	91.24	92.96
15	"suraj"	87.86	89.76	84.35	87.32
Average accuracy (in percentage) for words spoken In breathy voice					89.84

For breathy sound the overall accuracy obtained is 89.84% , but the error obtained in case of breathy is evenly distributed over the whole range of samples. Also for some words the nature of phoneme changes to unvoiced for *e.g.,* in case of "kalam" the /m/ phoneme becomes unvoiced which causes the accuracy rate for this word to decrease.

Finally Table 4.4 shows the overall accuracy of the algorithm for all types of voices.

Table 4.4: Overall accuracy (in percentage) obtained from algorithm.

Sr. No	**Voice Type**	**Accuracy (in percentage)**
1	Modal	91.05
2	Creaky	80.72
3	Breathy	89.84
Overall accuracy (in percentage)		87.2

The algorithm proposed is able to identify modal, creaky and breathy voices with an accuracy of 87.2%.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

After decades of work in the domain of speech recognition, voice quality still remains one the untouched part where very less work has been done and very less accuracy has been achieved. This thesis is an effort to improve the accuracy to identify different voice qualities.

The algorithm proposed in this work has successfully been able to identify the different voice qualities with some error rate especially in case of creaky voices. The algorithm was tested for 2 male speakers speaking in three different voice qualities (modal, breathy and creaky). For modal voice accuracy of 91.05% was achieved, for creaky it was a bit less ,*i.e.*, 80.72% and finally for breathy voice it was 89.84%. The overall accuracy of proposed algorithm is 87.2% which is quiet good considering the previous works.

The results achieved in present study motivate to extend the present work to achieve higher degree of accuracy especially in case of creaky voices where the algorithm shows less accuracy and also for breathy and modal voices more work can be done to achieve more accuracy in harsh environment and make the identification more robust.

REFERENCES

1) Atal, B., Rabiner, L., 1976. A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with applications to Speech Recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 2, no. 2, pp. 201-212.
2) Childers, D. G. and Ahn, C., 1995. Modeling the glottal volume-velocity waveform for three voice types. Journal of the Acoustical Society of America. vol. 97, no. 1, pp. 505-519.
3) Childers, D. G. and Lee, C.K., 1991. Vocal quality factors: analysis, synthesis, and perception. Journal of the Acoustical Society of America. vol. 90, no. 5, pp. 2394-2410.
4) Gerratt, B. R. and Kreiman, J., 2001. Towards a taxonomy of nonmodal phonation. Journal of Phonetics. vol. 29, pp. 365-381.
5) Gordon, M., 1998. The phonetics and phonology of non-modal vowels: a cross-linguistic perspective, Berkeley Linguistics Society. vol. 24, pp. 93-105.
6) Gordon, M., 2001. Linguistic aspects of voice quality with special reference to Athabaskan. Proceedings of the 2001 Athabaskan Languages Conference: 163-178.
7) Gordon, M and Ladefoged, P., 2001. Phonation types: a cross-linguistic overview, Journal of Phonetics, vol. 29, pp. 383-406.
8) Hillenbrand, J. and Cleveland R. A., 1994. Acoustic correlates of breathy vocal quality. Journal of Speech and Hearing Research. vol. 37, pp. 769-778.
9) Hillenbrand, J. and Houde, R. A., 1996. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. Journal of Phonetics, vol. 39, pp. 311-321.
10) Ishi, C. T., 2005. Analysis of autocorrelation-based parameters for creaky voice detection. Journal of acoustic science and technology, vol. 26, no. 4, pp.317-325.
11) Ishi, C. T., Sakakibara, K., Ishiguro and H., Hagita, N., 2008. A method for automatic detection of vocal fry. IEEE transactions on audio, speech, and language processing , vol. 16, no. 1, pp. 47-56.
12) Krom, G. D., 1993. A cepstrum based technique for determining a harmonic-to-noise ratio in speech signals. Journal of speech and hearing research, vol. 36, pp. 254-266.

13) Lee, J. Y., Jeong, S., Hahn, M., Choi, H. S., 2008. Automatic voice quality measurement based on efficient combination of multiple features. International conference on bioinformatics and biomedical engineering, vol. 4, pp. 2583-2586.
14) Malyska, N. and Quatieri, T. F., 2008. Spectral representations of nonmodal phonations. IEEE transactions on audio, speech and language processing, vol. 16, no. 1, pp. 34-46.
15) Mcloughlin, I., 2009. Applied Science and Audio Processing. Cambridge university press.
16) Rabiner, L. and Juang, B. H., 1993. Fundamental of Speech Recognition. PTR Prentice-Hall, New Jersey.
17) Rabiner, L. and Schafer, R. W., 2007. Introduction to Digital Speech Processing (Foundations and Trends in Signal Processing). Now Publications, Netherlands.
18) Shetye, A. S. and Carol Y. E., 2005. Analysis of modal and creaky voice quality variations. Journal of acoustic society of America, vol. 118, no. 3, pp. 1965-1965.
19) Shrivastav, R. and Sapienza, C. M., 2003. Objective measures of breathy voice quality obtained using an auditory model. Journal of the Acoustical Society of America, vol. 114, no. 4, pp. 2217-2224.
20) Wayland, R., Gargash, S., Longman, A., 1995. Acoustic and perceptual investigation of breathy voices. Journal of the Acoustical Society of America, vol. 97, no. 5, pp 3364-3364.
21) Yoon, T., Zhuang X. , Cole J. Johnson M., 2009. Voice quality dependent speech recognition. Linguistic Patterns in Spontaneous Speech (Language and Linguistics Monograph Series).
22) 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.