Acoustic Features Based Automatic Segmentation of Syllables

Thesis

submitted in partial fulfillment of the requirements

for the award of degree of

Masters of Technology

in

Computer Science and Applications

Submitted By

Harsimran Kaur

(Roll No. 601103006)

Supervised By

Dr. RK Sharma

School of Mathematics and Computer Applications

Thapar University

Patiala– 147004

July 2013

Certificate

I hereby certify that the work which is being submitted in the thesis entitled, **"ACOUSTIC FEATURES BASED AUTOMATIC SEGMENTATION OF SYLLABLES"**, 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.

Harsimran Kaur

(Harsimran Kaur)

Roll No. 601103006

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
15.7.13

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

Countersigned by:

15/07/2013

(Dr. Rajesh Kumar)
Head
School of Mathematics and Computer Applications (SMCA)
Thapar University
Patiala

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

Acknowledgments

It gives me immense pleasure to express my earnest gratitude to my supervisor, Dr. R.K. Sharma for his proficient guidance, persistent inspiration and encouragement. I am greatly thankful for his coherent support throughout the working of this dissertation. I feel extremely obliged to him for his comprehensive and thorough rectifications of the manuscript of this dissertation. I am indebted to Dr. Rajesh Kumar, Professor and Head, SMCA, Thapar University, Patiala for his continual support and cooperation.

I am grateful to the Almighty for the divine support and eternal motivation which have been steering me through the making up of this dissertation.

I am thankful to my parents who have always encouraged and supported me. Finally I appreciate the invaluable support and cooperation received from my family and friends.

Harsimran Kaur

(Harsimran Kaur)

Abstract

Automatic speech recognition (ASR) has intrigued researchers for the past several years and as such, significant contributions have been made in this field. The recognition process has been carried out at various levels, taking into consideration different speech units such as words, syllables and phonemes. The past few decades have witnessed substantial work in the field of automatic syllabification, *i.e.*, dividing a word into its constituent syllables, which in turn are comprised of phonemes. The pronunciation of a phoneme tends to vary depending on its location within a syllable. As such an acoustic analysis of phonemes has been carried out at a syllabic level. In order to achieve this goal, the automatic segmentation of a syllable into its constituent phonemes has been undertaken in the present work. In this work, the nasal consonants (ਮ /m/) and (ਨ /n/); and the vowels (ਅ /ə/, ਆ /a/, ਐ /æ/, ਏ /e/, ਈ /i/, ਓ /o/, ਔ /ɔ/ or ਊ /u/) in Punjabi language have been focused upon. Nasal are the only class of sounds that exhibit significant speech output from the nasal cavity as opposed to the oral cavity. Thus, it was of interest to examine how the nasal consonants may be perceived at a syllabic level. Putting into use the acoustic-phonetic approach to ASR, the acoustic features, namely, envelope variance, energy level and spectral peak frequency have been examined. It has been investigated as to what characteristics of these acoustic features are exhibited by the nasal consonants in context of the adjoining vowels. As such the focus of this study has been the syllables of the type: Nasal Consonant-Vowel (such as ਮਾ /maː/) and Vowel-Nasal Consonant (such as ਆਮ /aːm/). In order to carry out the automatic segmentation of these syllables, two approaches have been presented in this work. The first approach uses the change in the energy levels within a syllable to retrieve the point of segmentation of the syllable, while the second approach makes use of the change in the envelope variance of the syllable to perform segmentation. Further, the acoustic features have been used to train support vector machine (SVM) based classifier using LibSVM that in turn, identifies the nasal consonant part and the vowel part of the syllable.

The work undertaken in this thesis has been divided into six chapters. These chapters are: Introduction; Review of Literature; Data Collection, Preprocessing and Feature Extraction;

Acoustic Features based automatic segmentation of Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables; Results and Discussion; and Conclusion and Future Scope.

The first chapter describes ASR and its various approaches: acoustic-phonetic approach, pattern-recognition approach and artificial intelligence approach. Further, different manners of speech production and their correspondence to the vocal tract have been discussed. Since characteristics of the nasal consonants have been analyzed in context with the vowels, articulation and perception of the nasal consonants and the vowels have been discussed. The nasal consonants and the vowels of Punjabi language have been considered in the present thesis. Thus, a brief description of Punjabi phonology and Punjabi syllables has also been given in this chapter. Further, because of the significance of the spectral changes perceived in the nasal-vowel and vowel-nasal articulation, spectrograms of spoken syllables have been presented. A brief introduction of the SVMs has also been given in this chapter. In this work, SVM based classifiers have been trained using LibSVM in order to identify the nasal consonant part and the vowel part of a syllable.

The second chapter presents a survey of the research undertaken in the field of ASR. The review has been organized into three parts: the use of feature extraction techniques for automatic speech recognition, the research carried out for automatic segmentation of continuous speech; and the analysis of the nasals, semivowels and vowels.

In the third chapter, the various phases such as data collection, preprocessing and feature extraction have been discussed. In this study, collected data comprises of syllables consisting of Punjabi nasal consonants and vowels. The syllables collected are of two types: Nasal Consonant-Vowel and Vowel-Nasal Consonant. Since the recording has been done in noise-free environment, the preprocessing phase includes windowing and framing only. Further, the three acoustic features: envelope variance, energy level and spectral peak frequency have been explained in this chapter.

The fourth chapter elaborates the characteristics of the acoustic features introduced in chapter three. These features have been observed for the nasal consonants and the vowels considering them at a syllabic level. Further, the two approaches to automatic segmentation of syllables: using energy differences and using the envelope variance differences, have been described. The

algorithms developed for these approaches and the corresponding flowcharts have also been documented in this chapter.

In the fifth chapter, the results of the experiments that were carried out for automatic segmentation of Nasal Consonant-Vowel and the Vowel-Nasal Consonant syllables have been presented. The characteristics of the acoustical features: envelope variance, energy level and spectral peak frequency have been depicted graphically for the two syllable forms focused in this work. Further, recognition accuracy achieved in the experiments undertaken using LibSVM, for the two approaches, has been presented. Finally, a comparison of the two approaches to automatic segmentation of syllables has been documented. It has been observed that the automatic segmentation of the syllables (Nasal Consonant-Vowel and the Vowel-Nasal Consonant) gives better results when carried out using the envelope variance differences as compared to the approach that uses energy differences to perform the segmentation.

In the sixth chapter, conclusion, limitations and future scope of the present work have been presented in order to facilitate a refinement of this work.

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Expanded Form
ASR	Automatic Speech Recognition
AWGN	Additive White Gaussian Noise
CV	Consonant-Vowel
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
HMM	Hidden Markov Model
IPA	International Phonetic Alphabet
MRI	Magnetic Resonance Imaging
SVC	Support Vector Classifier
SVM	Support Vector Machine
VC	Vowel-Consonant
VNV	Vowel-Nasal-Vowel

Contents

Chapter 1

Introduction

Speech is the vocalized form of human communication. There has been a great desire that machines recognize human speech and respond accordingly. Embedding these capabilities in machines has intrigued researchers for many decades and as such major accomplishments have been achieved in this field. Depending upon the types of utterances, these have the ability to recognize, the ASR systems can be categorized as isolated word recognizer, connected word recognizer, continuous speech recognizer and spontaneous speech recognizer. ASR has been an important research activity in the previous years. The fundamentals of ASR have been described in this chapter.

1.1 Automatic Speech Recognition

ASR is the process of converting a speech signal to a sequence of words by means of an algorithm implemented as a computer program. A general model for speech recognition, shown in Figure 1.1 (Rabiner and Juang, 1993), depicts that the spoken output by a user is first decoded into a series of words that are meaningful according to the syntax, semantics and pragmatics of the recognition task. The meaning of the recognized words is obtained by a higher-level processor that uses a dynamic knowledge representation to modify the syntax, semantics and pragmatics according to the context of what it has previously recognized. The recognition system responds to the user in the form of the desired action.

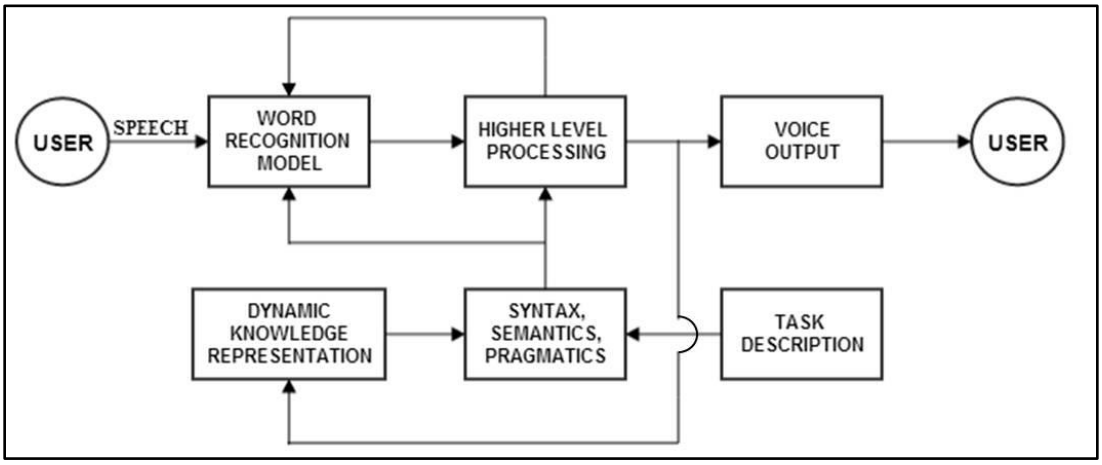

Figure 1.1: General block diagram of a task-oriented speech-recognition system

The various techniques to speech recognition can broadly be categorized as follows (Rabiner and Juang, 1993):

i. The acoustic-phonetic approach
ii. The pattern recognition approach
iii. The artificial intelligence approach

The acoustic phonetic approach to speech recognition relies on extracting a set of properties from a speech signal or its spectrum. Although the acoustic properties of phonetic units vary from speaker-to-speaker, it is assumed that the rules which govern the variability can be readily applied in practical situations. The second approach, namely, the pattern recognition approach involves pattern training and pattern comparison. This approach does not involve any feature extraction and segmentation. Instead the speech knowledge is brought into use through the training procedure and the patterns are recognized by the machine based on the training set provided to the algorithm. The artificial intelligence approach is a hybrid of the acoustic-phonetic approach and the pattern-recognition approach as it employs concepts of both methods. The tasks of segmentation and labeling are performed with more than just the acoustic information used by the acoustic-phonetic approach; using phonemic, lexical, syntactic and semantic knowledge.

The present study is a combination of the acoustic phonetic approach and the pattern recognition approach. Acoustic features have been extracted for some of the nasal consonants and the vowels of Punjabi language. Further using these acoustic features, SVM based classifier has been used to recognize these nasal consonants and vowels in the syllables of two types: Nasal Consonant-Vowel and Vowel-Nasal Consonant.

1.2 Manners of Speech Production

Speech involves successive narrowing and opening of the vocal tract, the passage through which the air flows during speech. The manner of articulation of a speech sound describes how the various parts of the vocal tract are involved in producing that sound. Table 1.1 enlists various manners of speech sounds and the corresponding shape of the vocal tract.

Table 1.1: Modes of Speech Production

Manner	Vocal Tract
Vowel	Open
Semivowel	Slightly constricted
Nasal	Closed (with nasal coupling)
Fricative	Narrow constriction
Stop	Completely closed

Of the modes of speech production enlisted in Table 1.1, this work carries out an analysis and recognition of the combination of nasal consonants and vowels at syllabic level.

1.2.1 Nasal Consonants

The articulation of nasals is characterized by the lowering of the velum that allows coupling to the nasal cavity and a complete closure in the oral cavity. The latter feature resembles the feature characterizing the production of stop consonants and the former gives the nasals their characteristic properties and reveals itself most prominently in the spectrum of the nasal murmur, that is, the sound produced with a complete closure at a point in the oral cavity, and with an appreciable amount of coupling of the nasal passages to the vocal tract. The articulatory system presented in Figure 1.2 (Fujimura, 1962) for the production of nasal murmurs consists of three subsystems.

1. The pharynx extending from the glottis to the velum
2. The oral cavity with a complete closure at the anterior end
3. The nasal tract including the nasopharynx and nasal passages that are terminated by radiation impedances

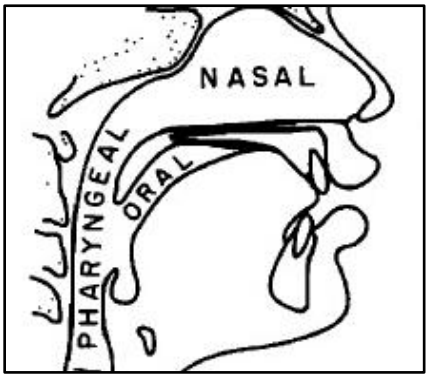

Figure 1.2: A tracing from a radiograph showing a midsaggital section of the vocal and nasal tracts during production of the nasal consonant /n/

The presence of a nasal consonant in a syllable can be characterized in the following three ways.

1. The transition between the nasal and an adjacent vowel where there is often an abrupt spectral change
2. The latter portion of a vowel preceding a nasal and the earlier portion of a vowel following a nasal where there might be nasalization because of the adjacent nasal
3. The region of the nasal murmur

1.2.2 Vowels

Vowels are sounds which occur at syllable centers because they involve a less narrowing of the vowel tract than consonants. A vowel is classified in terms of an abstract vowel space known as the vowel quadrilateral shown in Figure 1.3 (Handbook of the International Phonetic Association, 1999). This space bears a relation to the position of the tongue in vowel production as is evident from the mid-sagittal section of the vocal tract with four superimposed outlines of the tongue's shape depicted in Figure 1.4 (Handbook of the International Phonetic Association, 1999).

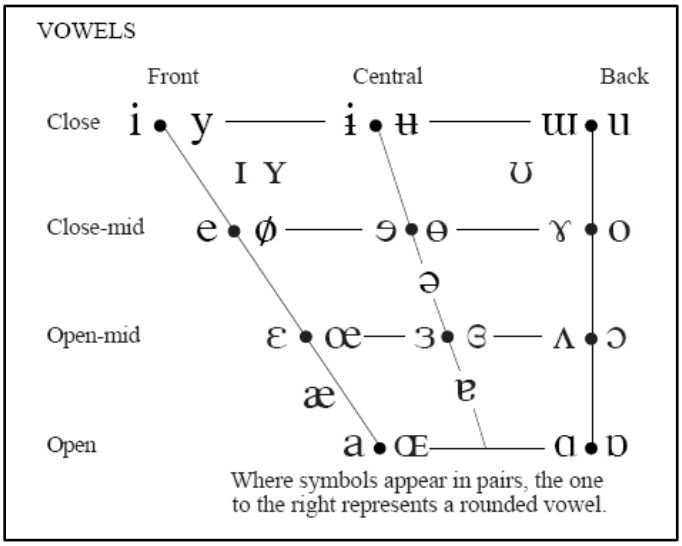

Figure 1.3: Vowel Quadrilateral

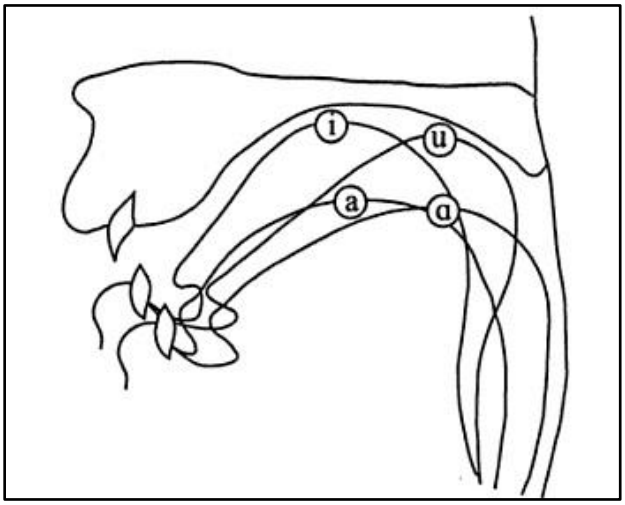

Figure 1.4: Mid-sagittal section of the vocal tract with the outline of the tongue shape for each of four extreme vowels imposed

The focus of the present work has been on syllables comprising of labial nasal consonant (ਮ /m/) and alveolar nasal consonant (ਨ /n/) and vowels (ਅ /ə/, ਆ /a/, ਐ /æ/, ਏ /e/, ਈ /i/, ਓ /o/, ਔ /ɔ/ or ਊ /u/) in Punjabi language. Syllables of the form Nasal-Vowel and Vowel-Nasal have been studied.

1.3 Punjabi Phonology

Punjabi, an Indo-Aryan language, is written using Gurmukhi and Shahmukhi scripts. In Gurmukhi script, which follows the *"one sound-one symbol"* principle, the Punjabi language has thirty eight consonants, ten non-nasal vowels and ten nasal vowels shown in Figure 1.5 (Arun, 1997).

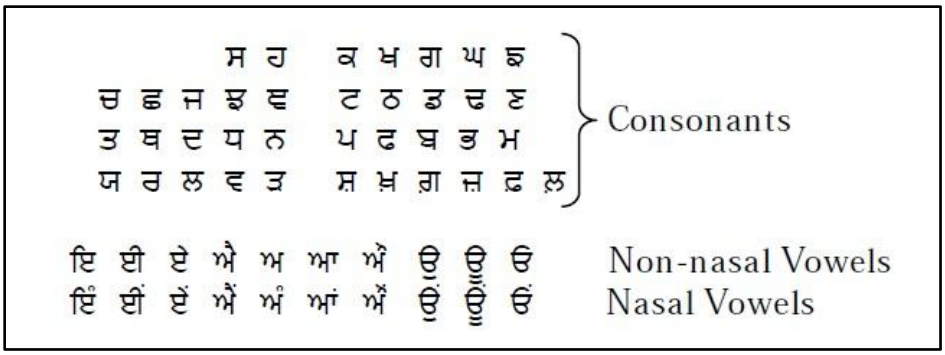

Figure 1.5: Punjabi Consonants and Vowels

In phonological approach, the syllables are defined by the different sequences of the phonemes. As such, combination of phonemes gives rise to next higher unit called syllable. Further combination of syllables produces larger units like morphemes and words. So syllable is a unit of sound which is larger than phoneme and smaller than word. In every language, certain sequences of phonemes and hence syllables are recognized. Phonologists are in general agreement that a syllable consists of a nucleus, preceded by an optional onset and followed by an optional coda. In Punjabi seven types of syllables are recognized; V, VC, CV, VCC, CVC, CVCC and CCVC (where V and C represents vowel and consonant respectively), which combine in turn to produce words.

1.4 Speech Signal Representation

Although some information about the phonetic content can be extracted from the waveform plots, yet in order to perform a detailed phonetic analysis, using short-time spectrum of the signal is more helpful. The short-time spectrum of the signal is the magnitude of a Fourier Transform of the waveform after it has been multiplied by a time window function of appropriate duration (Holmes and Holmes, 2001). An efficient representation of the speech signal based on short-time Fourier analysis is spectrograms. A spectrogram of a time signal is a special two-dimensional representation that displays time in the horizontal axis and frequency in the vertical axis. In order to indicate the energy in each time/frequency point, a grey scale is typically used, in which white represents low energy, and black, high energy (Huang *et al.*, 2001). Spectrograms can be also represented by a color scale where darkest blue parts represent low energy, and lightest red parts, high energy.

In the present study, a comprehensive analysis of the spectrogram of a syllable has been undertaken mainly because of the significance of the spectral changes perceived in the nasal-vowel and vowel-nasal articulation. Figure 1.6 and Figure 1.7 represent the waveforms as well as the spectrograms for VC syllable ਅਮ /am/ and CV syllable ਨੈ /næ/, respectively.

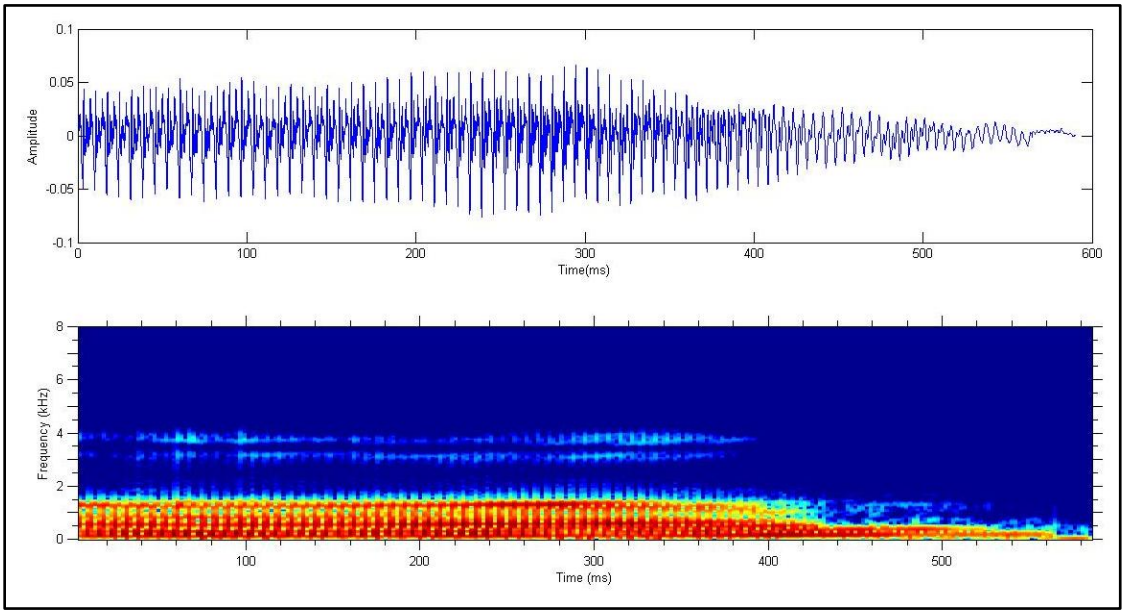

Figure 1.6: The waveform plot and the spectrogram of the Vowel-Nasal Consonant syllable ਅਮ

/am/

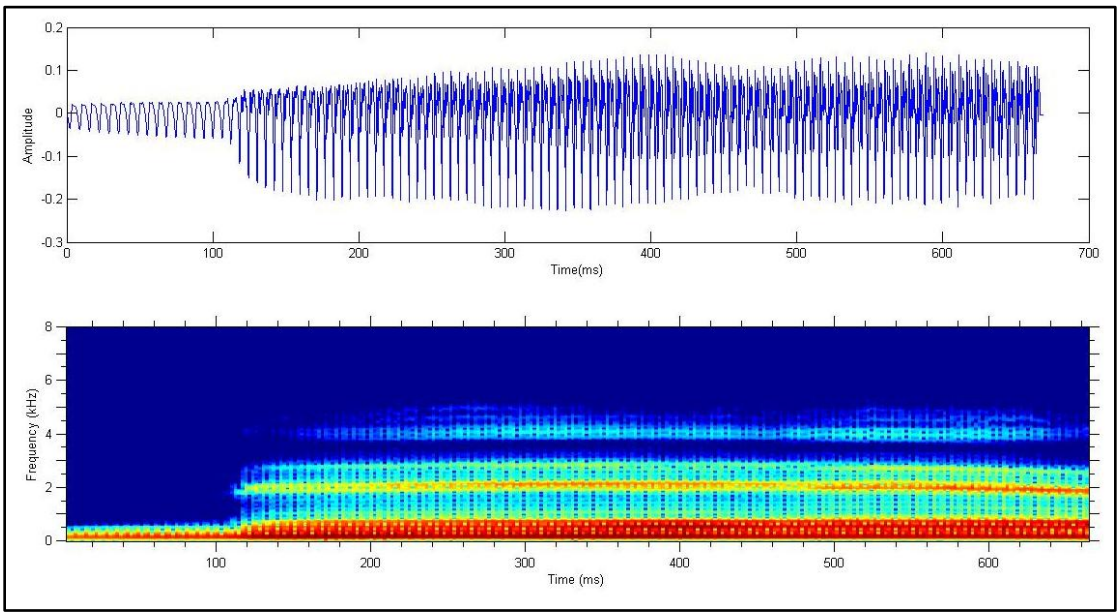

Figure 1.7: The waveform plot and the spectrogram of the Nasal Consonant-Vowel syllable ਨੈ

/næ/

1.5 Support Vector Machines (SVMs)

Support vector machines originally developed by Vladimir Vapnik in the 1990s, have a sound theoretical foundation rooted in statistical learning theory and require only as few as a dozen examples for training. For a two-class linearly separable learning task, the aim of support vector classifier (SVC) is to find a hyperplane that can separate two classes of given samples with a maximal margin which has been proved able to offer the best generalization ability. Generalization ability refers to the fact that a classifier not only has good classification performance on the training data, but also guarantees high predictive accuracy for the future data from the same distribution as the training data.

In the present study, once the Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables are automatically segmented, two SVCs have been trained using LibSVM in order to classify the segmented portions as belonging to class Nasal Consonant or class Vowel.

1.6 Problem Statement

Continuous speech often needs to be segmented into phonetic units for recognition and synthesis. As such, in the past decades, automatic segmentation of continuous speech has been developed to a great extent. Researchers have been successful in investigating and applying various techniques to carry out automatic segmentation up to the level of syllabification. The present thesis is an attempt to take this segmentation process a step further by examining phonemes by breaking a syllable down to the constituent phoneme-level. Since nasals are the only class of sounds that have dominant speech output from the nasal cavity rather than the oral cavity, it was of interest to examine how the nasal consonants are perceived in context of the adjoining vowels. Thus the problem at hand was to consider the syllables of the form: Nasal-Vowel and Vowel-Nasal and automatically segment the syllable into a nasal and a vowel; and further recognize the broken down phoneme as to whether it is a nasal or a vowel.

Chapter Summary

In this chapter, automatic speech recognition and its various approaches: acoustic-phonetic approach, pattern-recognition approach and artificial intelligence approach have been presented. Further, different manners of speech production and their correspondence to the vocal tract have been discussed. Of the various manners, nasals are the one that exhibit major speech output from the nasal tract as opposed to the oral tract. As such articulation of nasals and their perception have been documented. Since the nasal consonants have been analyzed in context with the vowels, articulation and perception of vowels have also been discussed. The performance of the task undertaken in the present thesis is evaluated using speech syllables from Punjabi language. Thus a brief description of Punjabi phonology and Punjabi syllables has been given in this chapter. Further, because of the significance of the spectral changes perceived in the nasal-vowel and vowel-nasal articulation, spectrograms of spoken syllables have been presented. SVMs have also been discussed in this chapter. The classification of nasal consonants and vowels has been carried out using two SVMs (each for Nasal Consonant-Vowel and Vowel-Nasal Consonant type syllables). Finally the problem statement of the current thesis has been documented in this chapter.

Chapter 2

Review of Literature

Research in ASR is being undertaken for almost five decades now and as such significant contribution by researchers from various disciplines: signal processing, pattern recognition, linguistics and computer science has been made. The recognition process has been carried out taking into consideration one of the speech units: phonetic features, phonemes, syllables, words or phrases. In each of the scenarios, a continuous speech signal has to be broken down to the desired speech unit and then analysis of that speech unit is performed. In the subsequent Section 2.1, the research carried out for the feature extraction based ASR has been documented. Section 2.2 presents a survey of the research pursued in the field of automatic segmentation of the speech signal. Further, Section 2.2 presents a review of the work done for analysis and recognition of the nasal consonants, semivowels and vowels at various syllabic and phonemic levels.

2.1 Feature extraction based ASR

ASR has been fascinating researchers and speech scientists for past several decades. As such, various approaches: the acoustic-phonetic approach, the pattern-recognition approach and the artificial intelligence approach have been put into use by the researchers.

In order to carry out a robust acoustic-phonetic analysis of the speech signal, extraction of linguistically relevant information has been proved helpful in the past years. The acoustic-phonetic analysis for speech recognition relies on modeling human knowledge of the speech signal. This involves extraction of features using signal processing tools and the decision making that interprets the extracted features in terms of speech units. Thus various areas concerned with human speech such as linguistics and phonetics get involved. The International Phonetic Association or the IPA is the oldest representative organization for phoneticians (Handbook of the International Phonetic Association, 1999). The main objective of this association is to promote a scientific study of phonetics. The IPA is the platform where speech signals can be studied irrespective of their languages. The IPA has formulated a set of symbols representing the wide variety of sounds found in the languages of the world. This set is referred to as the International Phonetic Alphabet or the IPA Chart. The IPA Chart is based on the Roman

alphabet, but has many additional symbols as well since the Roman set is not sufficient enough for all the languages of the world. This Chart is based on some assumptions listed below.

1. Some aspects of speech are linguistically relevant, while others (such as personal voice quality) are not.

2. Speech can be represented partly as a sequence of discrete sounds, called segments.

3. Segments are divided into consonants and vowels.

Being a vital component of the acoustic-phonetic approach to speech recognition, feature extraction has been put into use extensively to analyze the speech signal. Basically there are two approaches that classify various feature extraction techniques, temporal analysis and spectral analysis. In temporal analysis the speech waveform itself is used for analysis. On the other hand, in spectral analysis spectral representation of speech signal is used for analysis. Computing the energy levels of a speech signal has been carried out as part of the temporal analysis of the speech signal. In the traditional signal processing literature, the energy in a signal is considered to be the average of the sum of the squares of the magnitude of that signal. An alternate representation (Kaiser 1990) is to take the discrete Fourier transform (DFT) of that same signal segment where now the squares of the magnitudes of the frequency samples of the computed transform are assumed to represent the energy in the respective frequency components. This representation has been named Teager's energy operator. Jabloun *et al.* (1999) have proposed a new set of speech feature parameters that has been developed using multirate signal processing and the Teager's energy operator. It has been experimentally observed that the Teager's energy operator can suppress the car engine noise, which makes the new feature parameters a good candidate for voice dialing systems in automobiles.

Acoustic modeling and analysis of speech based on phonetic features has been explored in the work undertaken by Bitar (1998) for speaker-independent speech recognition. Phonetic features are minimal speech units that describe the manner and place of articulation of the sounds of a language. In this work, it has been shown that phonetic features have acoustic signatures in the speech signal that can be reliably extracted in a manner that reduces the effects of speaker differences. Algorithms that extract the acoustic properties of the phonetic features have been developed and these algorithms make measurements on the speech signal that are motivated by acoustic phonetics and spectrographic analysis. An event based recognition that uses these

measurements, combined by fuzzy rules has been developed and compared to a Hidden Markov Model (HMM) system using 1) the same measurements but modified to fit the frame-based HMM system and 2) Mel-cepstral parameters. The results show that the event based approach produces comparable results to the HMM frame-based system for the undertaken task of broad-class speech recognition. An automatic optimization procedure based on the Fisher criterion and classification trees has been developed to automate the derivation of acoustic measurements.

2.2 Automatic segmentation of continuous speech signal

For the purpose of recognition or synthesis, speech often needs to be segmented into phonetic units. Manual segmentation is tedious and time consuming and the results lack reproducibility because of the subjective decisions involved. This calls for automatic segmentation methods. Around five past decades have witnessed research in the field of segmentation of continuous speech.

Hemert (1990) proposed a method for automatic segmentation of speech used to create diphone libraries for Dutch, German and British English. Segmentation methods described in the literature can roughly be classified into two groups, implicit segmentation and explicit segmentation. Implicit segmentation methods split up the utterance into segments without explicit information, such as the phonetic transcription. Explicit segmentation methods split up the incoming utterance into segments that are explicitly defined by the phonetic transcription. In this study first an implicit segmentation algorithm splits up the utterance into segments on the basis of the degree of similarity between the frequency spectra of neighboring frames. Second an explicit algorithm does the same but this time on the basis of the degree of similarity between the frequency spectra of the frames in the utterance and reference spectra. A combined algorithm compares the two segmentation results and produces the final segmentation. Brugnara *et al.* (1993) have presented an automatic procedure for the segmentation of speech wherein given either the linguistic or the phonetic content of a speech utterance, phone boundaries are identified. An acoustic-phonetic unit Hidden Markov Model (HMM) recognizer has been used and both the recognizer and the segmentation system have been designed exploiting the DARPA-TIMIT acoustic-phonetic continuous speech database of American English. Putting neural networks into use, another system (Vorstermans *et al.*, 1996) has been developed for

automatic segmentation and labeling of speech originating from different languages without requiring extensive linguistic knowledge or large training databases of that language. Due to the limited size of the neural networks, the segmentation and labeling strategy requires but a limited amount of computations. The system was first evaluated on five isolated word corpora designed for the development of Dutch, French, American English, Spanish and Korean text-to-speech systems. Then additional tests were run on TIMIT and on the English, Danish and Italian portions of the EUROM0 continuous speech utterances.

Juneja and Espy-Wilson (2002) have presented a methodology for combining acoustic-phonetic knowledge with statistical learning for automatic segmentation and classification of continuous speech for recognition of broad classes; vowel, stop, fricative, sonorant consonant and silence. The technique makes use of 13 knowledge-based acoustic parameters (APs) and SVMs.
Bartlett *et al.* (2010) have presented several different approaches that segment words in English, Dutch and German. This study investigates approaches based on linguistic theories of syllabification as well as a discriminative learning technique that combines Support Vector Machine and Hidden Markov Model technologies. In this study three categorical approaches have been outlined based on common linguistic theories of syllabification, namely, the legality principle, the sonority sequencing principle and the maximal onset principle. Further it presents a Support Vector Machine Hidden Markov Model (SVM-HMM), which tags each phoneme with its syllabic role.

2.3 Analysis of nasal consonants, semivowels and vowels

The research undertaken for analysis and recognition of nasals (in various vowel contexts), semivowels and vowels has been presented. Researchers have developed various acoustic features for the nasals that have been investigated at a syllabic level. Apart from this, statistical models have also been used for their recognition.

An analysis-by-synthesis scheme (Fujimura, 1962) has been used to study the sound spectra of nasal murmurs in various vowel contexts. It has been inferred that the appearance of the spectra of nasal murmurs may vary from one another, depending on the individual nasal consonant and its context. The spectra may also depend on the individual speaker who utters the sound or on his

temporary physiological state as well. The study identifies the following acoustic characteristics that define nasal as a class.

1. The existence of a very low first formant that is located at about 300 Hz and is well separated from the upper formants structure
2. Relatively high damping factors of the formants
3. High density of the formants in the frequency domain

This study has focused on only one part of the acoustic continuum that is influenced by a nasal consonant, namely, nasal murmur. It is possible within the class of nasals to separate /m/, /n/ and /ŋ/ on the basis of the location of the antiformant, but the formant transitions of the adjacent vowels often play a more important role in the recognition of the individual nasals.

Kurowski and Blumstein (1987) focused on determining whether acoustic properties could be derived for English labial and alveolar nasal consonants that remain stable across vowel contexts, speakers and syllable positions. The study also investigated how the properties associated with nasal place of articulation relate to those proposed for articulation in stop consonants. It has been shown that for labial nasal consonants, there is a rapid increase in spectral energy in the lower frequency range relative to the higher range, and, for alveolar nasal consonants, there is a rapid increase in spectral energy in the higher frequency range relative to the lower frequency range. These acoustic properties have been generalized across three speakers in syllable-initial position and two speakers in syllable-medial position. Semivowels, being acoustically similar to the vowels, often occur adjacent to the vowels. This makes their identification more typical. Lately researchers have been investigating and working on the recognition of semivowels. In a feature-based approach to recognition, Espy-Wilson (1994) has developed a recognition system for the semivowels /j r l w/ in American English. The linguistic features studied for this work were sonorant, syllabic, consonantal, high, back, front and retroflex. Acoustic correlates and events related to these features were used to detect and classify the semivowels. Bitar and Espy-Wilson (1997) have designed a set of acoustic parameters (APs) that target phonetic features in the speech signal; sonorant, strident, syllabic, palatal, alveolar, labial and velar. The paper presents a two-stage procedure based on the Fisher criterion and automatic classification trees.

At times, nasal consonants are confused in the presence of background noise. Alwan *et al.* (1999) have studied the perception of the place of articulation for nasal consonants in adverse conditions examined through a series of perceptual experiments. The experiments examined the effects of additive white Gaussian noise (AWGN), and additive speech-shaped noise on nasal place perception in CV syllables. A Hidden Markov Model (HMM)-based ASR system has been constructed to identify the nasals at various signal-to-noise ratios.

When carrying out the task of segmentation of continuous speech, the landmarks in a continuous speech signal need to be identified. Salamon *et al.* (2004) presented the use of temporal information for extraction of linguistically relevant details from a speech signal. In this study, a system is developed that extracts linguistically relevant temporal information that can be used in the front end of an automatic speech recognition system. The parameters used are energy onset and offsets and measures of periodic and aperiodic content. These have been combined to find abrupt acoustic events which signify landmarks. This work has also shown that there may be certain landmark types where spectral features and perhaps more subtle temporal features (on a longer time scale) are important, particularly for landmarks related to sonorant consonants.

Further nasal consonants have been compared with semivowels by Pruthi and Espy-Wilson(2004) who have focused on the development of Acoustic Parameters (APs) which can be extracted automatically and reliably in a speaker independent way. These APs have been tested in a classification experiment between nasals and semivowels, the two classes of sounds which together form the class of sonorant consonants. In this study, the following four APs have been identified.

1. An energy onset/offset measure to capture the consonantal nature of nasals, that occurs at the closure and release of nasal consonants
2. An energy ratio
3. A low spectral peak measure to capture the nasal murmur
4. An envelope variance parameter to capture the spectral and temporal stability during the nasal murmur

Ramanarayanan *et al.* (2010) have presented a comparative study focusing on the details of the nasal consonant in read and spontaneous speaking styles. In this study, vowel-nasal-vowel (VNV) segments from one speaker have been examined and the focus has been on investigating

the acoustic spectral reduction of read and spontaneous speech production differences. The acoustic spectral reduction is characterized by the reduction in spectral and durational distinctions between sounds as the speaking style becomes more informal or the stress on the syllable is reduced.

Recently Vuppala *et al.* (2012) have proposed an efficient approach to spotting and recognition of consonant-vowel (CV) units from continuous speech using accurate detection of vowel onset points (VOPs). The proposed method is carried out at two levels consisting of hidden Markov models (HMMs) at the first stage for recognizing the vowel category of a CV unit and support vector machines (SVMs) for recognizing the consonant category of a CV unit at the second stage.

Although various aspects of phoneme recognition have been investigated by researchers, there is a lot of scope for developing a set of time domain techniques that involve the wave form of the speech signal directly. Sunil Kumar and Lajish (2012) have studied the variation of average energies in the zero crossing interval of the speech signal and have evaluated the distribution of parameters throughout the signal. It has been observed that the distribution patterns are almost similar for repeated utterances of the same vowels and vary from vowel to vowel. This distribution pattern is used for recognizing five Malayalam vowels using multilayer feed forward artificial neural network.

Thus the review of the research done so far has revealed that automatic segmentation of continuous speech into syllables, that is, syllabification has been examined by various researchers in the past. Whereas automatic segmentation of the syllable into phonemes has not been explored much in the past especially using the feature-based approach. Thus the current thesis is an attempt to study the acoustic phonetic features of nasal consonants and vowels and further put these features into use in order to automatically break down a syllable comprising of these two phonemes.

Chapter Summary

In this chapter a survey of the research undertaken in the field of automatic speech recognition has been documented. Further, various approaches using statistical models, acoustic features as well as neural networks for the automation of the segmentation process have been presented. This segmentation of the speech signal has been carried out taking into consideration various speech units such as syllables and phonemes. The research for the analysis and recognition of nasal consonants, semivowels and vowels at phonemic and syllabic levels has been presented in this chapter.

Chapter 3

Data Collection, Preprocessing and Feature Extraction

In order to carry out an analysis of a speech signal, it has to go through various phases, namely, data collection, preprocessing and feature extraction. These three phases have been discussed in context with the present analysis of syllables in subsequent sections.

3.1 Data Collection

Since the identification of the nasal consonants and the vowels has been performed at a syllable level, the data set for this study comprises of syllables in two forms; Nasal Consonant-Vowel (such as ಮಾ /maː/) and Vowel-Nasal Consonant (such as ಆಮ್ /aːm/). The syllables have been recorded by 3 female speakers and 3 male speakers using the tool COLEA. The syllable samples have been recorded with a sampling frequency of 16000 Hz and stored using 16 bits in a wave format. The recording has been carried out in a normal room environment.

3.2 Preprocessing

Preprocessing means carrying out various operations on the speech signal in order that the useful part can be extracted from it. This part can be subjected to desired acoustical analysis. The preprocessing stage may involve all or some of the processes such as noise removal, pre-emphasis, framing and windowing, depending on the nature of the input signal. For the present study, preprocessing steps that have been included are framing and windowing because the input data samples have been recorded in a normal noise-free room environment. These processes have been discussed in subsequent subsections.

Framing and Windowing

Segmentation of the speech signal into frames is a basic requirement of audio processing as extracting features becomes complex for a larger speech signal. This process of segmentation has its own problems. When an audio feature, that is, a vector of samples characterizing some type of sound, is segmented into two frames, it is quite possible that one half of that audio feature appears in one audio frame and the other half in the second frame. As a result, the actual

properties of that feature are not retrieved. To overcome this problem, the speech vector is segmented into overlapping frames. This ensures that the audio features occurring at a discontinuity are considered as a whole in the subsequent overlapped frame. The process of framing the speech signal involves application of a window function to the speech vector. In the present work, Hanning window has been applied and this is defined below.

Hanning Window

The Hanning window is named after Julius von Hann and is also known as the von Hann window and the raised cosine window. It is defined (Heinzel *et al.*, 2002) as,

$$w_j = 0.5\left(1 - cos\left(\frac{2\pi j}{N-1}\right)\right)$$

where $j = 0 \ldots N\text{-}1$; N is the length of the discrete Fourier transform (DFT).

A Hanning window is used by taking j number of sample points and thus is referred to as j-point Hanning window. A 50-point Hanning window is shown in Figure 3.1.

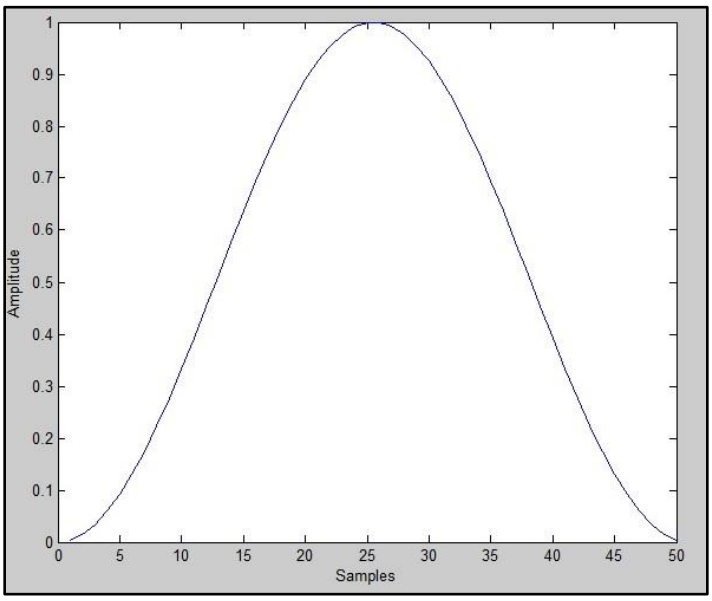

Figure 3.1: Hanning Window Function

3.3 Feature Extraction

Analyzing and extracting relevant information from the speech signal is a crucial task as this is what determines the efficiency of the recognition systems. The selection of those features that exhibit similar properties across various speakers and are able to characterize a particular mode of speech such as nasals and vowels is an important step of the recognition process. Basically there are two approaches that classify various feature extraction techniques, temporal analysis and spectral analysis. In temporal analysis the speech waveform itself is used for analysis. Temporal features are easy to extract and have a simple physical interpretation. On the other hand, in spectral analysis spectral representation of speech signal is used for analysis. The time domain data is converted to frequency domain by applying Fourier transform to it. The main structure of a feature extractor is shown in Figure 3.2.

For the present work, two spectral features, envelope variance and spectral peak frequency (Pruthi and Espy-Wilson, 2004) and one temporal feature, energy levels have been used. The acoustic features extracted to identify nasal consonants and vowels in Punjabi language are described in subsequent sections.

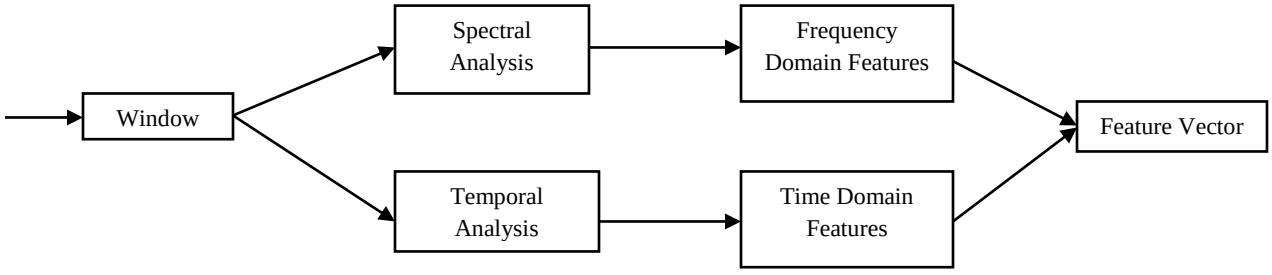

Figure 3.2: Feature Extractor

Acoustic Features

Acoustic features are measures performed on the speech waveform and its time- frequency transformations in order to seek evidence for the acoustic properties of phonetic features. In order to analyze the nasal consonants and the vowels in different syllable contexts, the acoustic features that have been focused upon in this thesis are envelope variance parameter to capture the spectral and temporal stability, the energy levels and spectral peak frequency.

3.3.1 Envelope Variance

The envelope function of a rapidly varying signal is a smooth curve outlining its extremes in amplitude. The envelope function may be a function of time, space, angle or any variable. An envelope is interpreted as the absolute value of a complex representation of a signal. As such Hilbert transform has been applied to compute the envelope of the syllables.

Hilbert Transform

The Hilbert transform (Hahn, 1996) $H[g(t)]$ of a signal$g(t)$, where *t* stands for time, is defined as,

$$H[g(t)] = g(t) * \frac{1}{\pi t} = 1/\pi \int_{-\infty}^{\infty} \frac{g(\tau)}{t-\tau} * d\tau = \int_{-\infty}^{\infty} \frac{g(t-\tau)}{\tau} * d\tau$$

The Hilbert transform of $g(t)$ is the convolution of $g(t)$ with the signal $1/\pi t$. Thus for the present study, the Hilbert envelope has been computed by taking the Hilbert transform of the signal and then considering the absolute value of the transformed signal.

3.3.2 Teager Energy

In the traditional signal processing literature, the energy in a signal is considered to be the average of the sum of the squares of the magnitude of that signal. Another representation of energy of a signal known as Teager's energy operator (Kaiser, 1990) is being used in various applications. It is defined as,

$$\mathrm{E}[x_i(n)] = \sum_{1}^{k} [x_i(n) * x_i(n) - x_i(n-1) * x_i(n+1)]$$

where *k* is the number of speech samples in the i^{th} frame; $x_i(n)$ is the amplitude of the n^{th} speech sample of the i^{th} frame.

In the present study, the energy levels for each frame have been calculated using the Teager's energy operator.

3.3.3 Spectral Peak Frequency

The frequency spectrum of a time-domain signal is a representation of that signal in the frequency domain and it can be generated through a Fourier transform of the signal. The Fourier transform of a function produces a frequency spectrum which contains all of the information about the original signal, but in a different form. Frequency analysis or spectrum analysis can either be performed on the entire signal or the signal can be broken into frames and spectrum analysis may be applied to each individual frame. Figure 3.3 shows the signal waveform and the spectrogram of the Punjabi word ਵਰਦੀ (uniform).

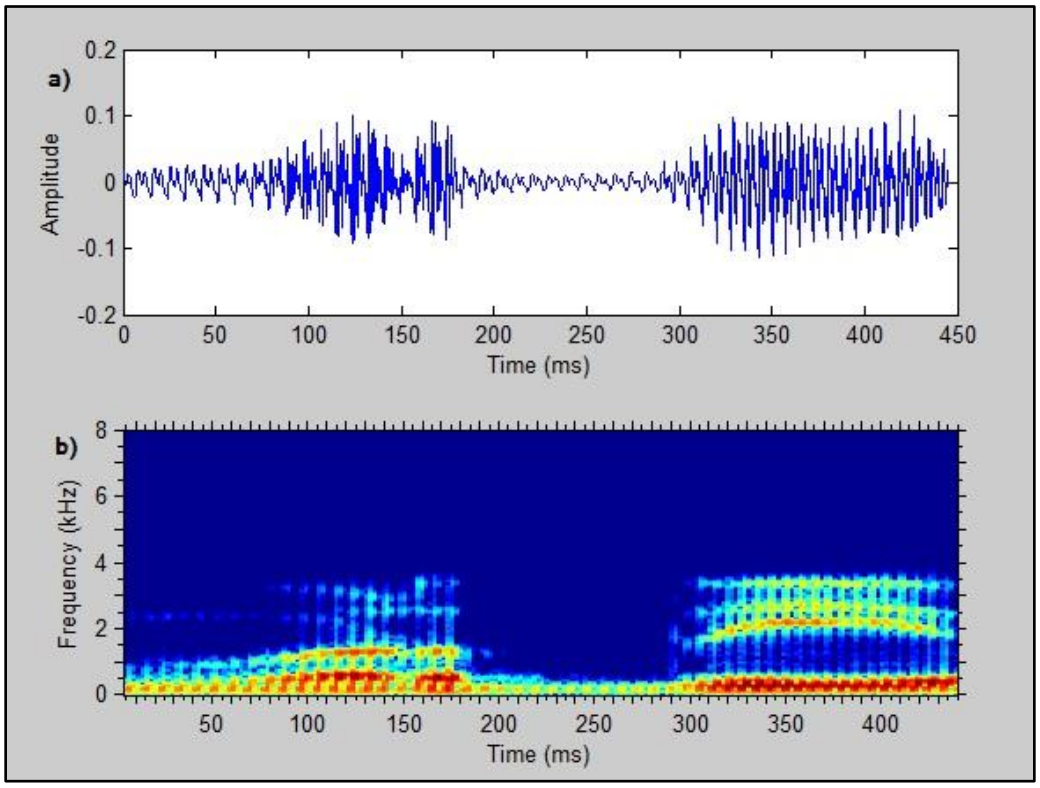

Figure 3.3: a) Signal Waveform b) Spectrogram of the signal

For the current thesis, the spectral peak frequency measure calculated is the frequency corresponding to the maximum fast Fourier transform (FFT) spectrum in 0-800 Hz range. The frequency range used in this work has been defined in (Bitar and Espy-Wilson, 1997).

3.4 Statistics on data collection

In order to carry out the automatic segmentation of syllables in Punjabi of the form: Nasal Consonant-Vowel and Vowel-Nasal Consonant, speech samples have been recorded by native speakers of Punjab. The analysis and recognition has been done for syllables comprising of the nasal consonants (/m/ and /n/) and vowels (/ə/, /a/, /æ/, /e/, /i/, /o/, /ɔ/ or /u/). The focus has been on two types of syllables: Nasal Consonant-Vowel and Vowel-Nasal Consonant. The statistics data collected for the two experiments conducted in this work is shown in Table 3.1 and Table3.2.

Table 3.1: Statistics for Nasal Consonant-Vowel syllables

Speaker ID	Gender	Age	No. of Samples
1	Female	50	50
2	Male	26	48
3	Male	24	85
4	Male	12	18

Table 3.2: Statistics for the Vowel-Nasal Consonant syllables

Speaker ID	Gender	Age	No. of Samples
1	Female	50	56
2	Female	35	30
3	Female	26	20
4	Male	26	38
5	Male	24	30
6	Male	12	28

Chapter Summary

The analysis of a speech signal requires it to undergo certain preprocessing procedures. In the present chapter framing and windowing applied on the Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables have been discussed. In order to recognize the nasal consonants and vowels, three features, namely, envelope variance, Teager energy and spectral peak frequency have been described. Further the details of the data samples recorded for the experiments have been documented.

Chapter 4

Acoustic Features based automatic segmentation of Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables

The analysis and evaluation of the three acoustic features have revealed certain characteristics of the occurrence of nasal consonants and vowels in a syllable. These properties have proved to be contributive in order to perform the automatic segmentation of the syllable (Nasal Consonant-Vowel as well as Vowel-Nasal Consonant). Further, these have been useful in identifying the two segmented parts of the syllable as nasal consonant and vowel. The characteristics of the acoustic features, namely, envelope variance, Teager energy and spectral peak frequency for both nasal consonants as well as vowels have been discussed in subsequent sections.

4.1 Characteristics of the acoustic features perceived at Nasal Consonant-Vowel and Vowel-Nasal Consonant syllable level

The syllables of Punjabi language, comprising of the nasal consonants and the vowels, have been observed and as such, some interesting characteristics of the acoustic features have been perceived. These have been discussed below.

4.1.1 Characteristics of the envelope variance

It has been observed that the envelope of a nasal consonant remains constant as compared to that of a vowel. This is evident from the plots shown in the Figure 4.1 and Figure 4.2 of the syllables {ਨਾ} /naː/ and {ਆਨ} /aːn/ consisting of a nasal consonant and a vowel (both Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables). Figure 4.1(b) and Figure 4.2(b) represent the Hilbert envelope for both the syllables {ਨਾ} /naː/ and {ਆਨ} /aːn/. The standard deviation value of the Hilbert envelope is calculated for both the nasal consonant and the vowel subparts of the syllable. The nasal consonant part of the syllable exhibits a low standard deviation value as compared to that of the vowel.

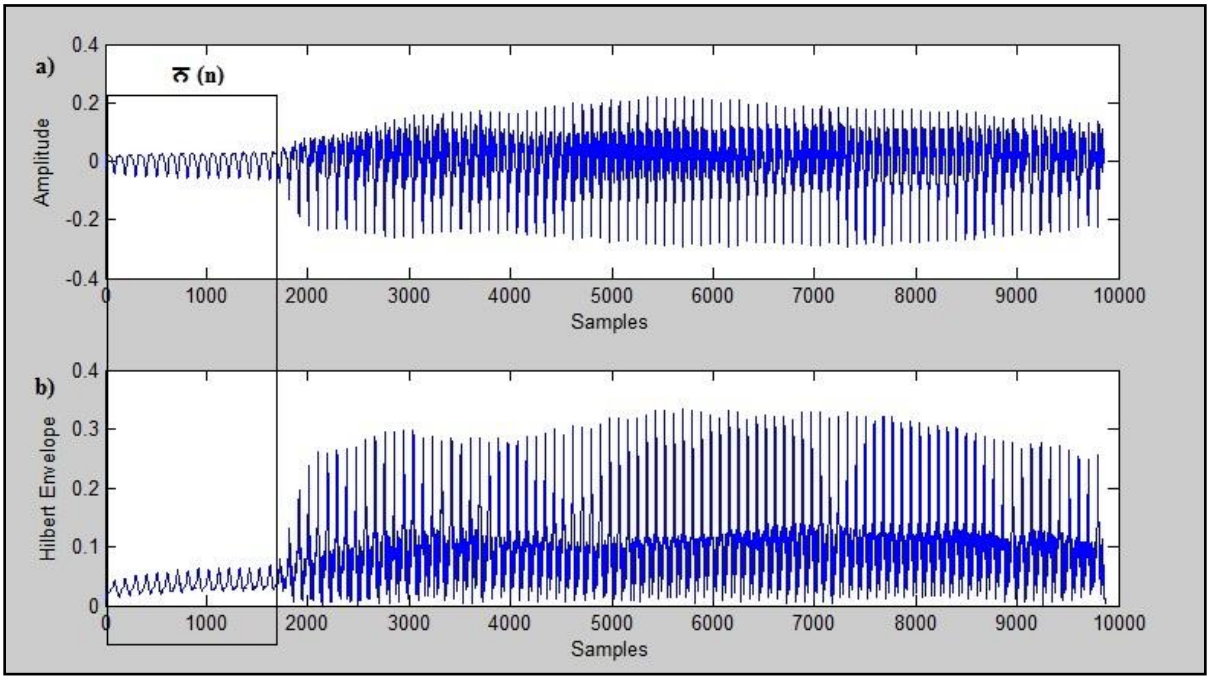

Figure 4.1: Nasal Consonant-Vowel Syllable ਨਾ /naː/ a) Signal Waveform b) Hilbert Envelope

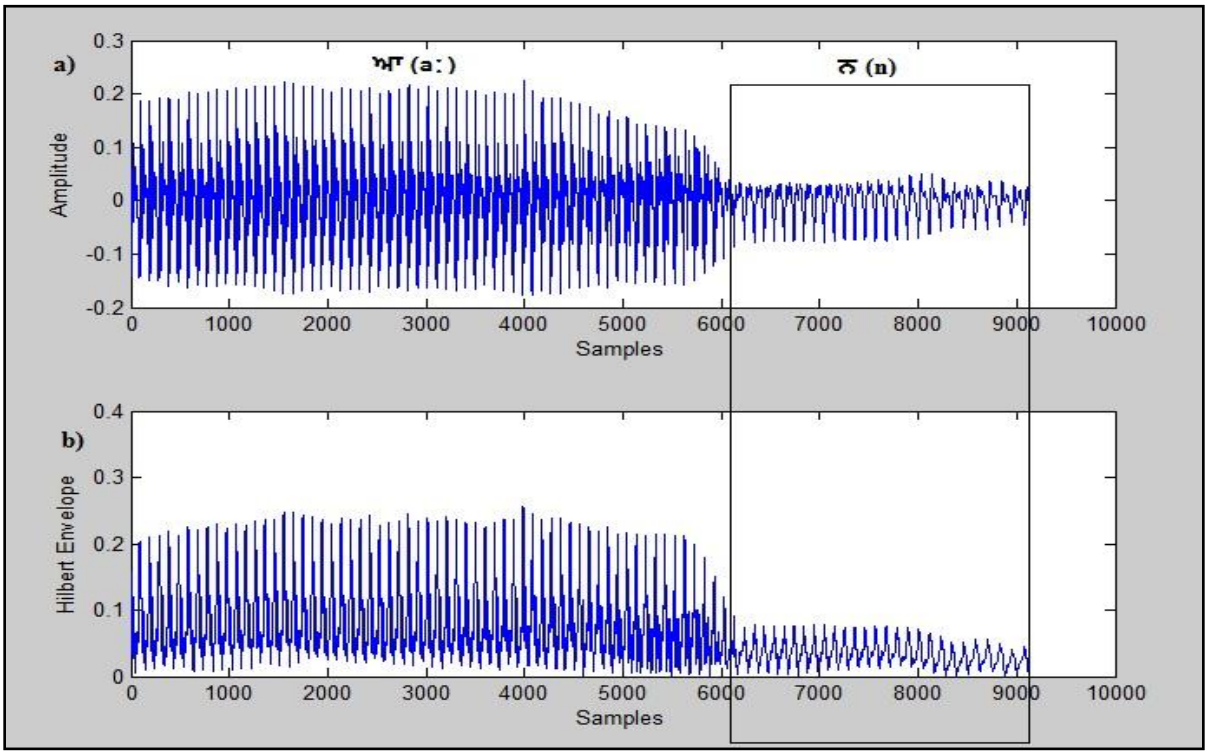

Figure 4.2: Vowel-Nasal Consonant Syllable ਆਨ /aːn/ a) Signal Waveform b) Hilbert Envelope

4.1.2 Characteristics of the signal's energy levels

In a syllable consisting of a nasal consonant and a vowel, there is sharp energy onset and energy offset at nasal consonant-vowel and vowel-nasal consonant boundaries respectively. The Teager energy levels for both the nasal consonant and the vowel parts of the syllables {ਨਾ} /naː/ and {ਆਨ} /aːn/ are shown in Figure 4.3(b) and Figure 4.4(b). Thus it can be noticed that the vowels tend to have higher energy values as compared to the nasals.

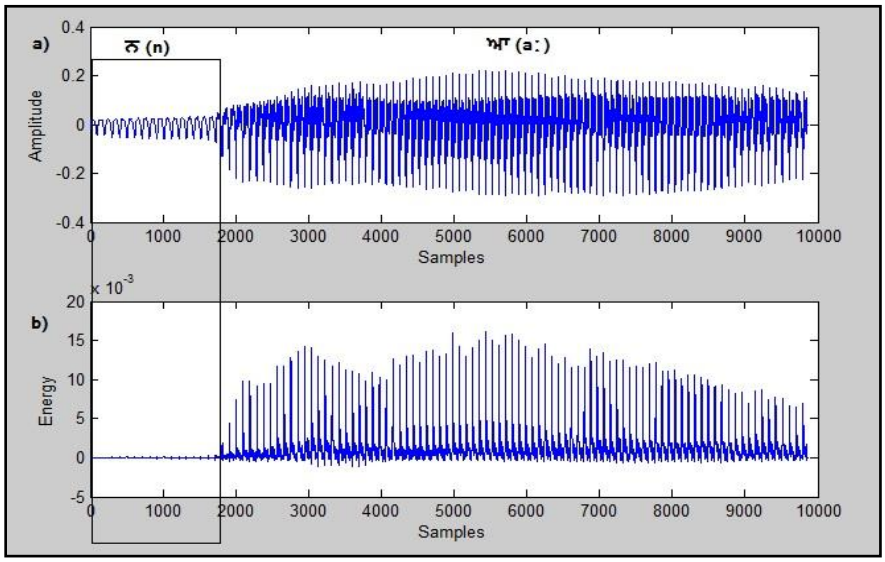

Figure 4.3: Nasal Consonant-Vowel Syllable ਨਾ /naː/ a) Signal Waveform b) Energy of the signal

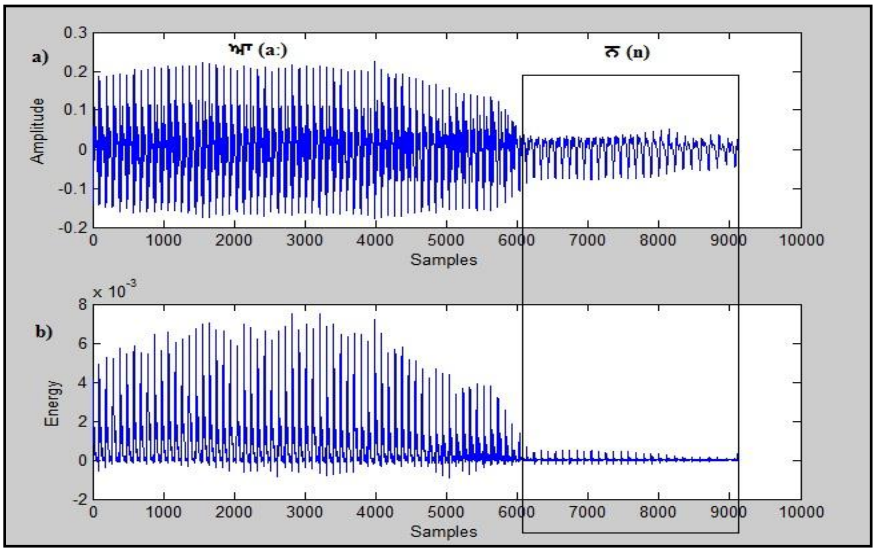

Figure 4.4: Vowel-Nasal Consonant Syllable ਆਨ /aːn/ a) Signal Waveform b) Energy of the signal

4.1.3 Characteristics of the spectral peak frequency

Nasals tend to have a lower frequency value for the first spectral prominence as compared to vowels. In case of the Nasal consonant-Vowel syllable ਨਾ /naa/, the spectrogram in the Figure 4.5(b) shows that spectral prominence occurs at about 450 Hz for the nasal consonant and at about 1500 Hz for the vowel. Similarly for the Vowel-Nasal consonant syllable ਆਨ /a:n/, the spectrogram in the Figure 4.6(b) shows that spectral prominence occurs at about 450 Hz for the nasal consonant and at about 1000 Hz for the vowel.

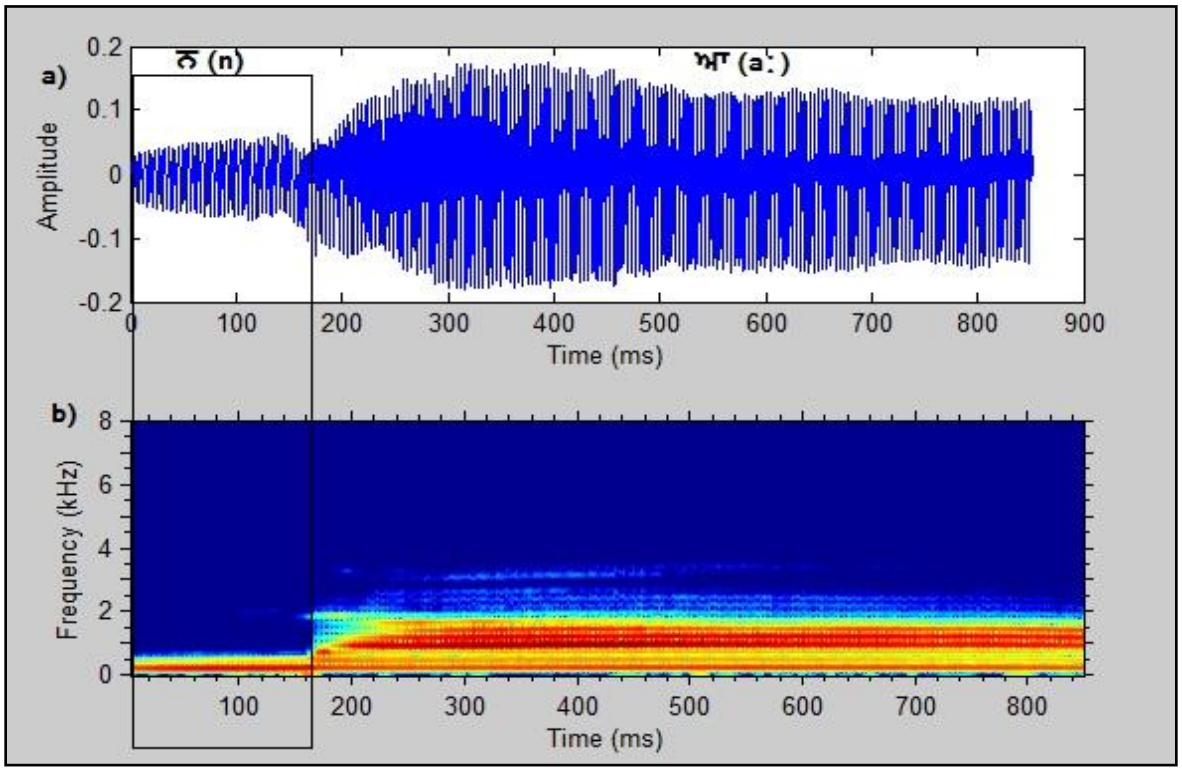

Figure 4.5: Nasal Consonant-Vowel Syllable ਨਾ /na:/ a) Signal Waveform b) Spectrogram of the signal

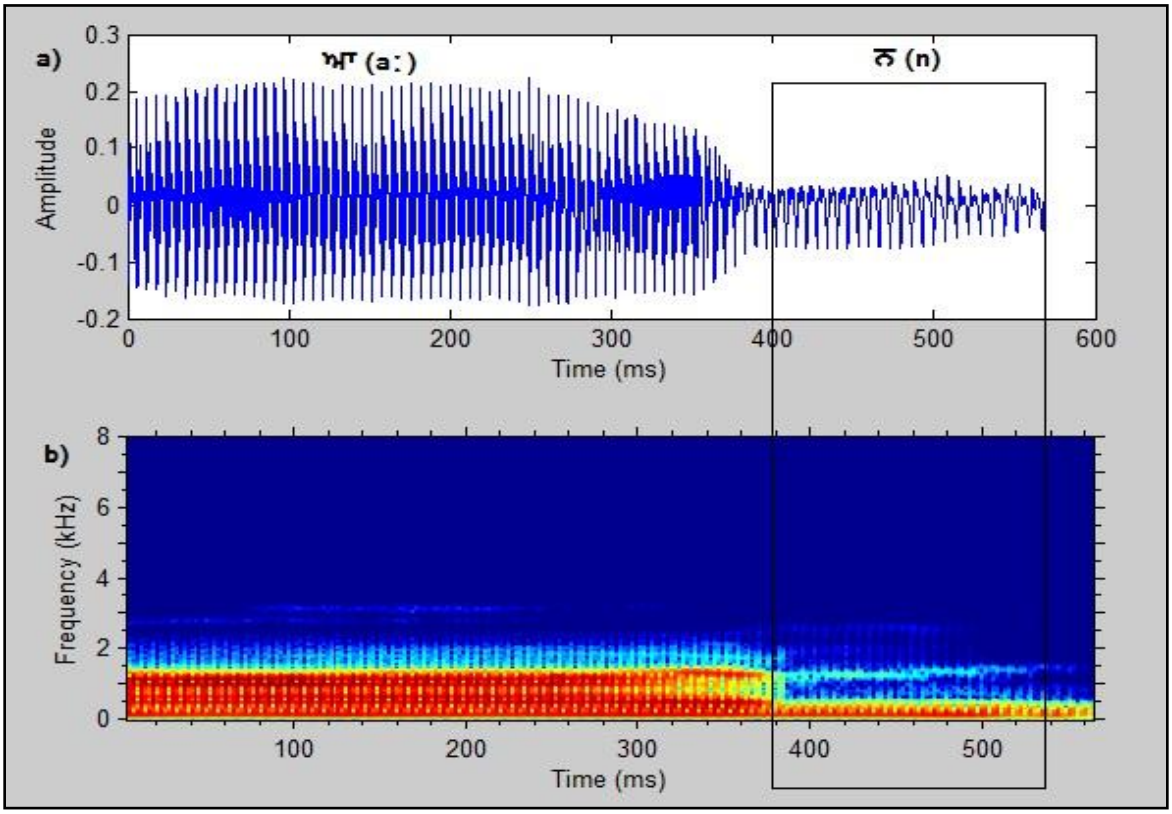

Figure 4.6: Vowel-Nasal Consonant Syllable ਆਨ /aːn/ a) Signal Waveform b) Spectrogram of the signal

4.2 Automatic Segmentation of the Nasal consonant-Vowel and the Vowel-Nasal consonant syllables

For the purpose of recognition, speech often needs to be segmented into phonetic units. Manual segmentation is tedious and time consuming, and the results lack reproducibility because of the human knowledge involved. This calls for automatic segmentation methods. The present work deals with automatic segmentation of syllables consisting of Punjabi nasal consonants and vowels. In order to identify the nasal consonant and the vowel, the syllable needs to be segmented into phonemes. For this purpose, two segmentation techniques have been presented and compared. These are described in subsequent subsections.

4.2.1 Automatic Segmentation of the syllable using the energy differences

The first method that has been applied to carry out segmentation of the Nasal consonant-Vowel and the Vowel-Nasal consonant syllables puts into use the energy differences observed within a syllable. The degree of abruptness in the energy differences measure has been used to identify the point of segmentation of the syllable. After preprocessing, the speech signal is segmented into frames and the energy levels for each frame are computed. The energy of the speech chunk within each frame is calculated using the Teager energy operator (Kaiser, 1990), which is defined as follows,

$$\mathrm{E}[x_i(n)] = \sum_{1}^{k} [x_i(n) * x_i(n) - x_i(n-1) * x_i(n+1)]$$

where *k* is the number of speech samples in the i^{th} frame; $x_i(n)$ is the amplitude of the n^{th} speech sample of the i^{th} frame.

Then energy differences for each frame are calculated using the following formula,

$$Energy\_Diff(i) = E[x_i(n+1)] - E[x_i(n)]$$

where *n* varies from 1 to *k*; *k* being the number of audio samples in the i^{th} frame.

The average energy change in each frame is computed as follows,

$$Avg\_Energy(i) = Energy\_Diff(i)/NOF$$

where *i* represents the frame index and *NOF* stands for Number of Frames.

Once we get the average energy change for all the frames, the frame with the maximum energy change is chosen as the point of segmentation of the syllable into two subparts. Thereafter the acoustic features described in earlier sections, namely, Teager energy, standard deviation of the Hilbert envelope and spectral peak frequency measure are computed for both the subparts (nasal consonant and vowel). The analysis of nasal consonants and vowels has shown that the energy level of a nasal consonant is much lower than the energy level of a vowel.

An algorithm has been developed to automatically segment the Nasal Consonant-Vowel type syllables using the energy differences. The Algorithm 4.1 has been presented below.

Algorithm 4.1: Automatic Segmentation of Nasal Consonant-Vowel syllables using energy differences

i. Read the sound file and create a speech vector.
ii. Partition the speech vector into overlapping frames using a Hanning window of size 20 ms with an overlap of 5 ms.
iii. Compute the energy levels vector (E) for each frame.
iv. Compute the energy differences vector (ED) for each frame.
v. Calculate the average energy change for each frame.
vi. Retrieve the frame with the maximum average energy change and take the index of that frame as the point of segmentation of the syllable into two subparts.
vii. Compute energy vectors (EN) for the frames in the first and the second subparts.
viii. Calculate the average energy values for both the subparts.
ix. Compute the Hilbert envelopes for the frames in the first and the second subparts.
x. Compute the standard deviation (SD) of the Hilbert envelopes for the frames in the first and the second subparts.
xi. Calculate the average standard deviation values for both the subparts.
xii. Compute the Spectral Peak Frequency vectors (SPF) for the frames in the first and the second subparts.
xiii. Calculate the average Spectral Peak Frequency values for both the subparts.
xiv. Store the three average values (Energy, Standard Deviation of Hilbert envelope and Spectral Peak Frequency) in libsvm format in a text file with label 1 for the first subpart and with label 2 for the second subpart.

A flowchart depicting the steps of the above algorithm has been shown in Figure 4.7.

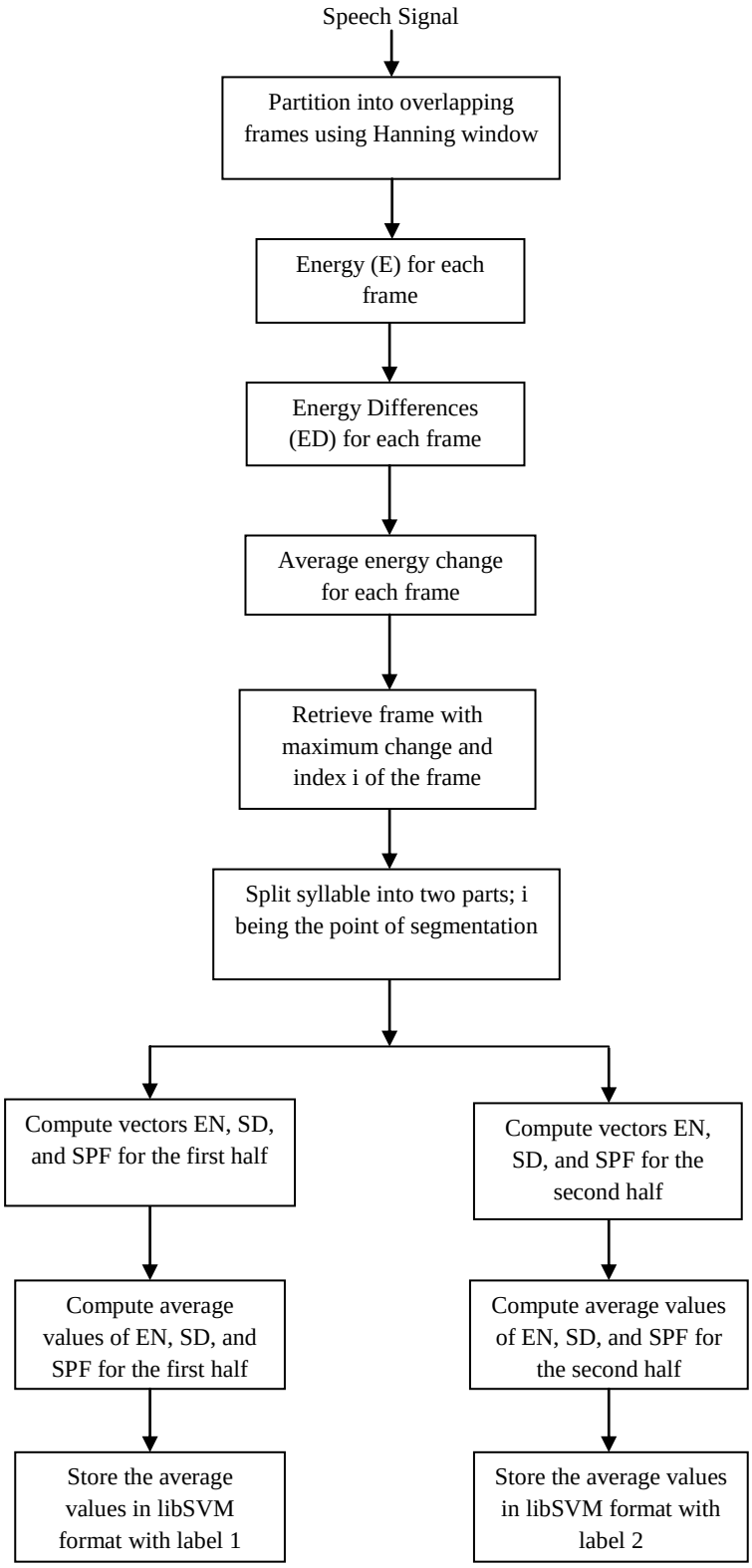

Figure 4.7: Flowchart for automatic segmentation of Nasal Consonant-Vowel syllables using energy differences

The Algorithm 4.1 has been applied to 200 Punjabi (Nasal Consonant-Vowel) syllables consisting of nasal consonants (/m/ or /n/) and vowels (/ə/, /a/, /æ/, /e/, /i/, /o/, /ɔ/ or /u/) recorded by 4 (3 male and 1 female) speakers.

Secondly, the automatic segmentation using energy differences has been implemented on Vowel-Nasal Consonants of Punjabi language. An algorithm (Algorithm 4.2) developed for this technique has been documented below.

Algorithm 4.2: Automatic Segmentation of Vowel-Nasal Consonant syllables using energy differences

i. Read the sound file and create a speech vector.
ii. Partition the speech vector into overlapping frames using a Hanning window of size 20 ms with an overlap of 5 ms.
iii. Compute the energy levels vector (E) for each frame.
iv. Compute the energy differences vector (ED) for each frame.
v. Calculate the average energy change for each frame.
vi. Retrieve the frame with the maximum average energy change and take the index of that frame as the point of segmentation of the syllable into two subparts.
vii. Compute energy vectors (EN) for the frames in the first and the second subparts.
viii. Calculate the average energy values for both the subparts.
ix. Compute the Hilbert envelopes for the frames in the first and the second subparts.
x. Compute the standard deviation (SD) of the Hilbert envelopes for the frames in the first and the second subparts.
xi. Calculate the average standard deviation values for both the subparts.
xii. Compute the Spectral Peak Frequency vectors (SPF) for the frames in the first and the second subparts.
xiii. Calculate the average Spectral Peak Frequency values for both the subparts.
xiv. Store the three average values (Energy, Standard Deviation of Hilbert envelope and Spectral Peak Frequency) in libSVM format in a text file with label 2 for the first subpart and with label 1 for the second subpart.

The flowchart for the above algorithm is presented in Figure 4.8.

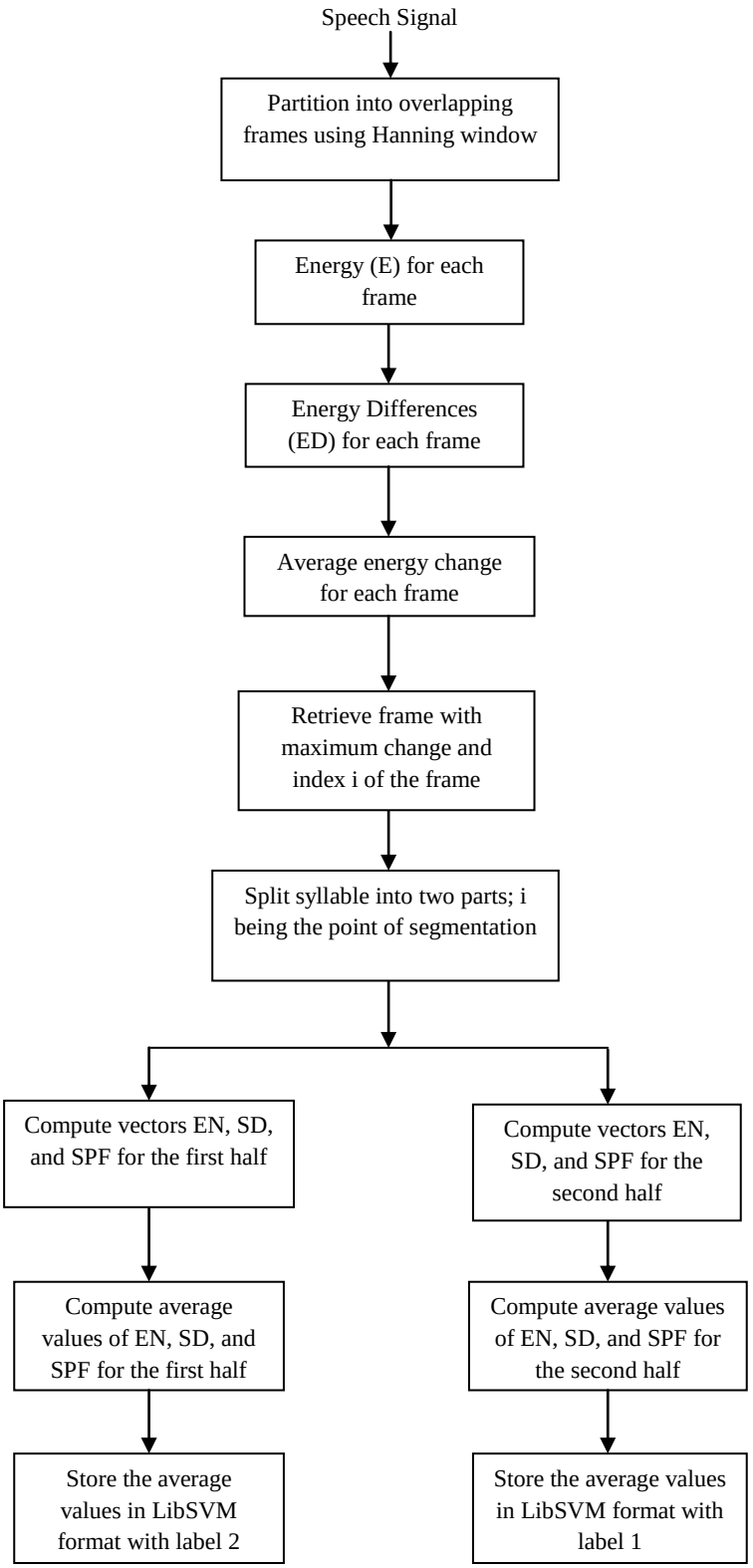

Figure 4.8: Flowchart for automatic segmentation of Vowel-Nasal Consonant syllables using energy differences

The Algorithm 4.2 has been applied to 200 Punjabi (Vowel-Nasal Consonant) syllables consisting of nasal consonants (/m/ or /n/) and vowels (/ə/, /a/, /æ/, /e/, /i/, /o/, /ɔ/ or /u/) recorded by 6 (3 male and 3 female) speakers.

4.2.2 Automatic Segmentation of the syllable using the envelope variance differences

In the previous section, application of differences in the energy levels has been employed to partition the syllable. In addition to that, another technique has been described to attain the same goal. It has been observed that nasal consonants tend to have much more constant envelope as compared to vowels. In order to perform segmentation, the Hilbert envelope for each frame of the speech signal is computed and then standard deviation of the Hilbert envelope is calculated. The vector $Std(i)$ represents standard deviation measure for i_{th} frame. The differences in the standard deviation are computed as,

$$Std\_Diff(i) = Std(i+1) - Std(i)$$

where *i* represents the frame index.

Once the vector *Std_Diff* is computed, the frame with the maximum difference is chosen to be the frame where the segmentation is carried out and the syllable is partitioned into two subparts. Thereafter the three acoustic parameters: Teager energy, standard deviation of the Hilbert envelope and spectral peak frequency measure are computed for both the subparts (nasal consonant and vowel).

An algorithm has been developed to perform the bisection of the Nasal consonant-Vowel syllables and further evaluate the values of the acoustic features to identify the nasal consonant and the vowel subsections. This Algorithm 4.3 has been presented below.

Algorithm 4.3: Automatic Segmentation of Nasal Consonant-Vowel syllables using change in envelope variance

i. Read the sound file and create a speech vector.
ii. Partition the speech vector into overlapping frames using a Hanning window of size 20 ms with an overlap of 5 ms.
iii. Compute Hilbert envelope for each frame.
iv. Calculate standard deviation (S) of the Hilbert envelope for each frame.
v. Find the differences in the standard deviation (SD_Diff) measure of the frames.
vi. Retrieve the frame with the maximum standard deviation difference and select the index of that frame as the point of bisection of the syllable into two subsections.
vii. Compute energy vectors (EN) for the frames in the first and the second subsection.
viii. Calculate the average energy values for both the subsections.
ix. Compute the Hilbert envelopes for the frames in the first and the second subsections.
x. Compute the standard deviation (SD) of the Hilbert envelopes for the frames in the first and the second subsections.
xi. Calculate the average standard deviation values for both the subsections.
xii. Compute the Spectral Peak Frequency vectors (SPF) for the frames in the first and the second subsections.
xiii. Calculate the average Spectral Peak Frequency values for both the subsections.
xiv. Store the three average values (Energy, Standard Deviation of Hilbert envelope and Spectral Peak Frequency) in Libsvm format in a text file with label 1 for the first subsection and with label 2 for the second subsection.

An outline of the above algorithm has been depicted in the form of a flowchart as shown in Figure 4.9.

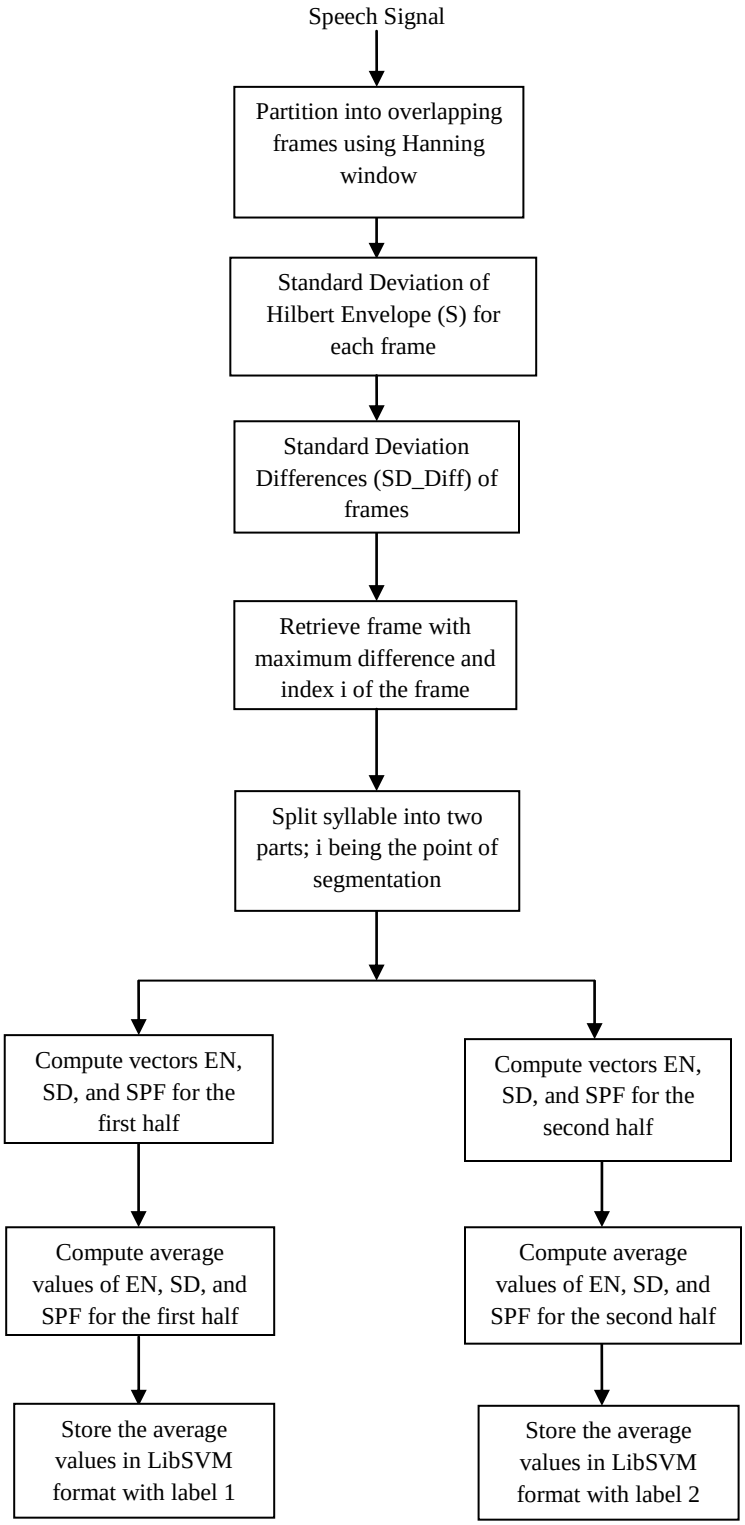

Figure 4.9: Flowchart for automatic segmentation of Nasal Consonant-Vowel syllables using change in envelope variance

The Algorithm 4.3 has been implemented on 200 syllable samples, of Nasal Consonant-Vowel type, recorded by 4 (3 male and 1 female) speakers of Punjabi language.

The implementation of the use of envelope variance for segmentation purposes has also been applied to the Vowel-Nasal Consonant type syllables in Punjabi language. The Algorithm 4.4 provides the details of this implementation technique.

Algorithm 4.4: Automatic Segmentation of Vowel-Nasal Consonant syllables using change in envelope variance

i. Read the sound file and create a speech vector.
ii. Partition the speech vector into overlapping frames using a Hanning window of size 20 ms with an overlap of 5 ms.
iii. Compute Hilbert envelope for each frame.
iv. Calculate standard deviation (S) of the Hilbert envelope for each frame.
v. Find the differences in the standard deviation (SD_Diff) measure of the frames.
vi. Retrieve the frame with the maximum standard deviation difference and select the index of that frame as the point of bisection of the syllable into two subsections.
vii. Compute energy vectors (EN) for the frames in the first and the second subsection.
viii. Calculate the average energy values for both the subsections.
ix. Compute the Hilbert envelopes for the frames in the first and the second subsections.
x. Compute the standard deviation (SD) of the Hilbert envelopes for the frames in the first and the second subsections.
xi. Calculate the average standard deviation values for both the subsections.
xii. Compute the Spectral Peak Frequency vectors (SPF) for the frames in the first and the second subsections.
xiii. Calculate the average Spectral Peak Frequency values for both the subsections.
xiv. Store the three average values (Energy, Standard Deviation of Hilbert envelope and Spectral Peak Frequency) in LibSVM format in a text file with label 2 for the first subsection and with label 1 for the second subsection.

A flowchart exhibiting the steps described in the above algorithm has been presented in Figure 4.10.

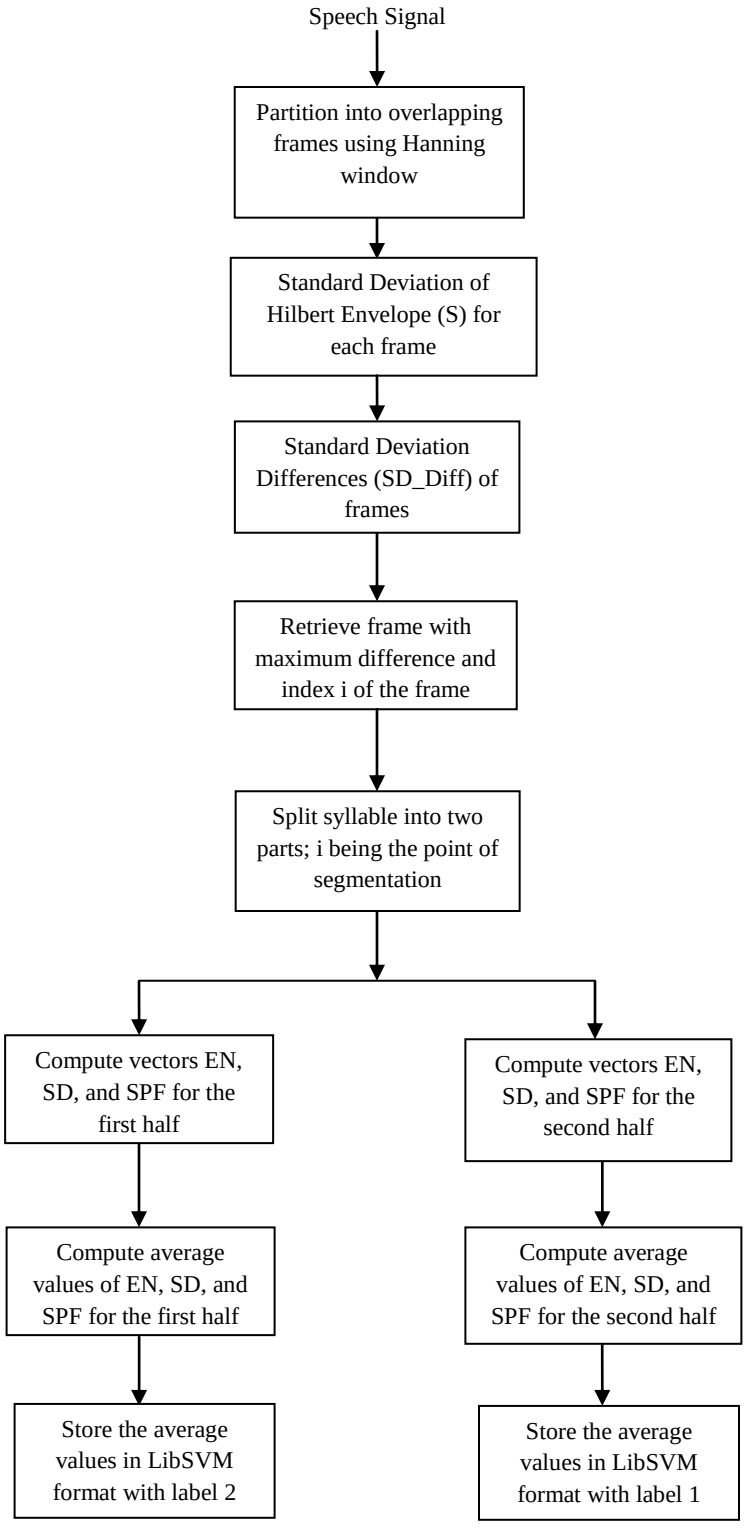

Figure 4.10: Flowchart for automatic segmentation of Vowel-Nasal Consonant syllables using change in envelope variance

The Algorithm 4.4 has been applied to 200 Vowel-Nasal Consonant syllable samples spoken in Punjabi language by 6 (3 male and 3 female) speakers. The syllables consist of Punjabi nasal consonants (/m/ or /n/) and vowels (/ə/, /a/, /æ/, /e/, /i/, /o/, /ɔ/ or /u/) such as {ਆਨ} /aːn/.

The experiments that have been carried out for classification and the corresponding results have been documented in the next chapter.

Chapter Summary

In this chapter, an analysis of the three acoustic features, namely, envelope variance, Teager energy and spectral peak frequency has been documented. The characteristics of these features for Punjabi nasal consonants and vowels at a syllable level have been presented. Further, the chapter presents the algorithms and flowcharts for two different techniques: temporal and spectral that can be used to perform segmentation of the syllable. Once the Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables are bisected, the values for the three features are extracted for both nasal consonants and vowels in order to recognize them in a syllable.

Chapter 5

Results and Discussion

The two approaches for automatic segmentation of the syllables, discussed in the previous chapter, have been implemented in MATLAB using the speech processing toolbox, named, Voicebox. Once the values of the acoustic features are retrieved, these have been used to train SVM based classifier using the software LibSVM, that in turn, identifies the nasal consonant and the vowel in a syllable. The experiments carried out for the two approaches and the corresponding results have been discussed in subsequent sections.

5.1 Experimentation for data analysis

It is evident from the discussions in the previous chapters that the nasal consonants tend to have lower values for acoustic features (envelope variance, energy level and spectral peak frequency) as compared to those of the vowels. In case of the Nasal Consonant-Vowel and the Vowel-Nasal Consonant syllables, an analysis has been carried out, taking into consideration, 50 data samples spoken by 4 native speakers (2 male and 2 female) of Punjabi language. The graphs shown in figures (Figure 5.1, Figure 5.2 and Figure 5.3) for the Vowel-Nasal Consonant syllables and the graphs in figures (Figure 5.4, Figure 5.5 and Figure 5.6) for the Nasal Consonant-Vowel syllables provide ample evidence of the characteristics of the acoustic features for nasal consonants and vowels.

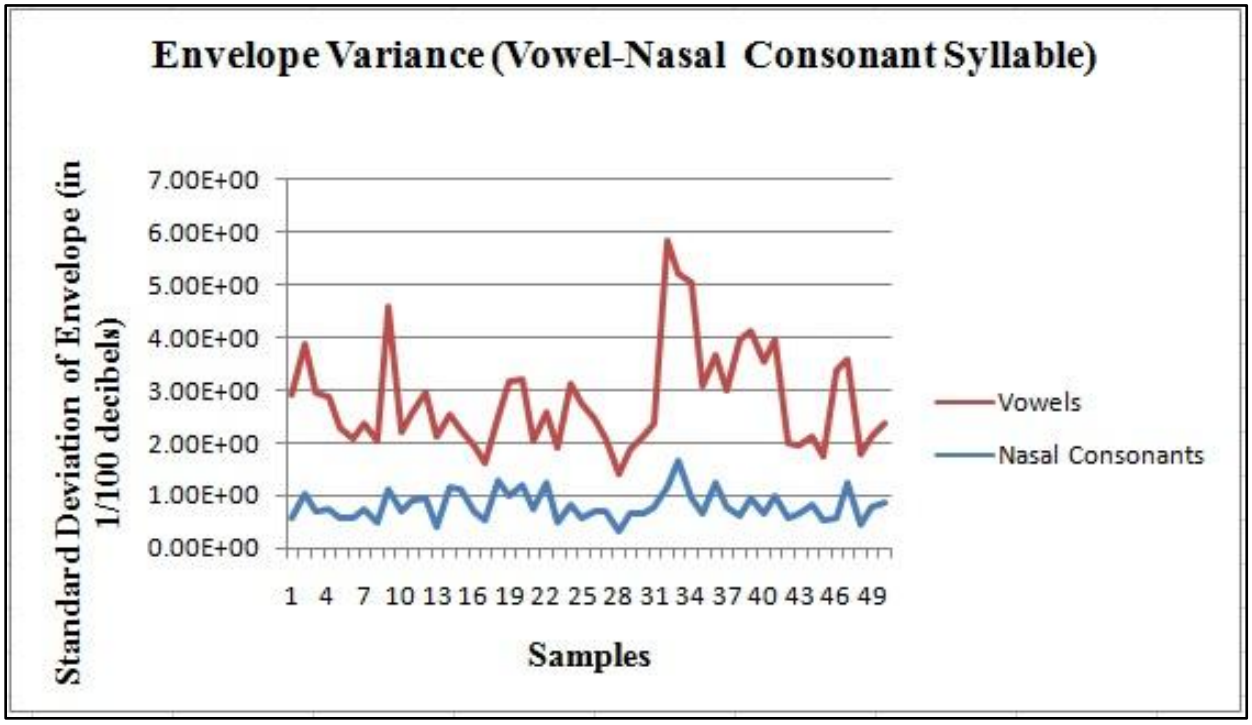

Figure 5.1: Graphs for the envelope variance of Vowel-Nasal Consonant syllables

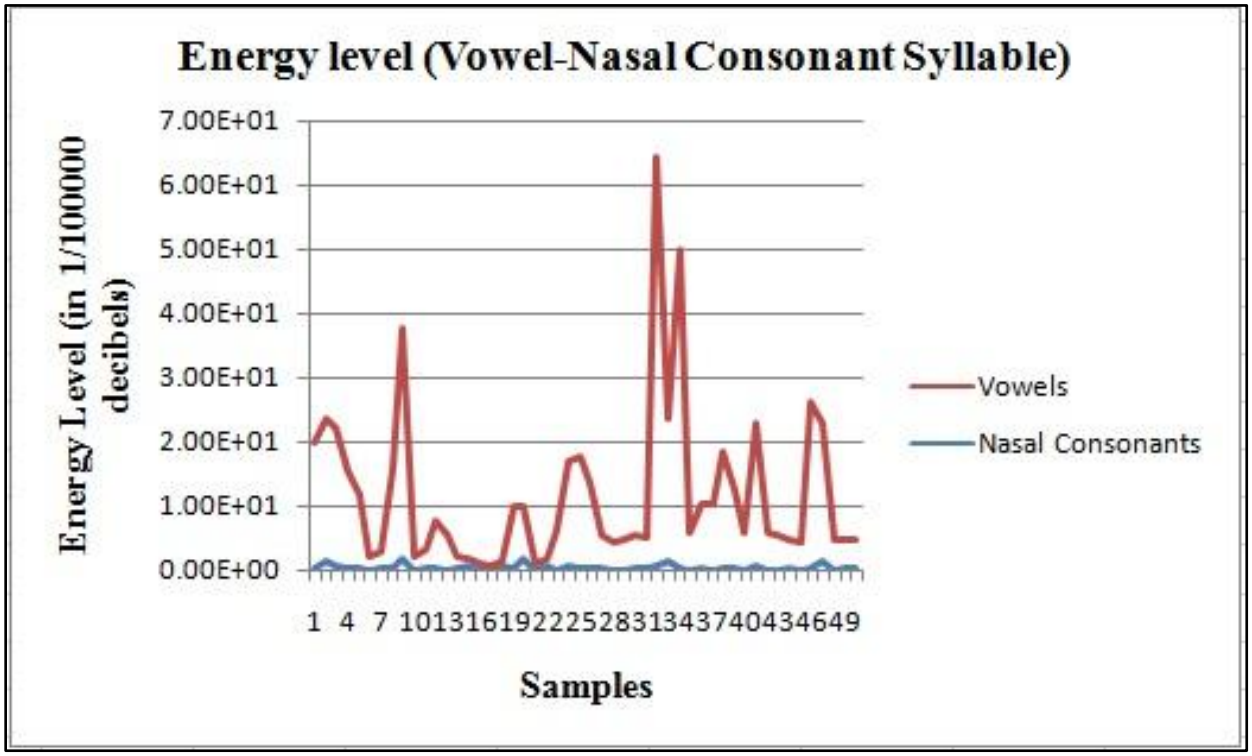

Figure 5.2: Graphs for the energy levels of Vowel-Nasal Consonant syllables

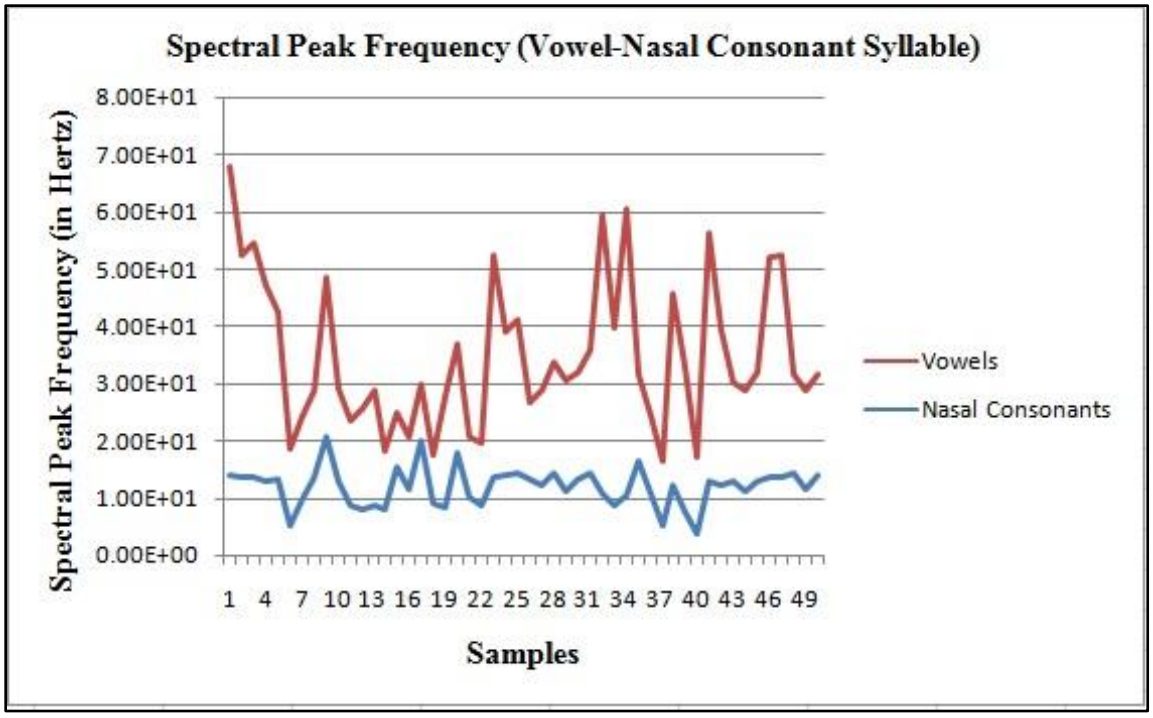

Figure 5.3: Graphs for the spectral peak frequency of Vowel-Nasal Consonant syllables

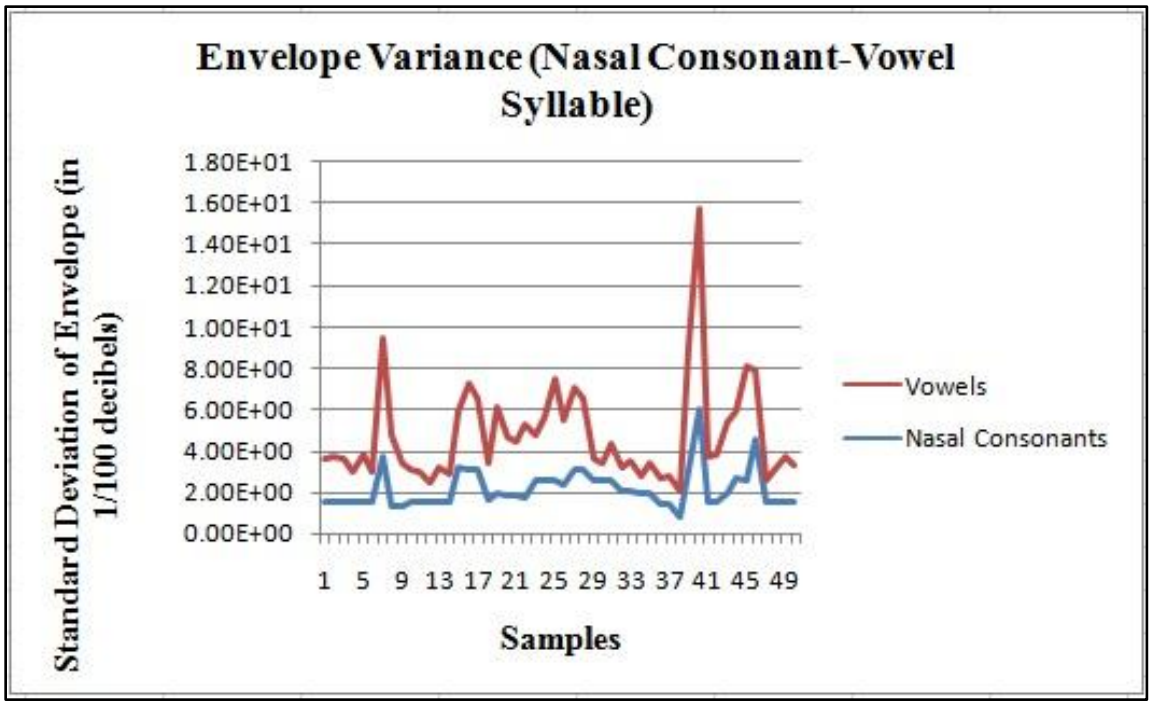

Figure 5.4: Graphs for the envelope variance of Nasal Consonant-Vowel syllables

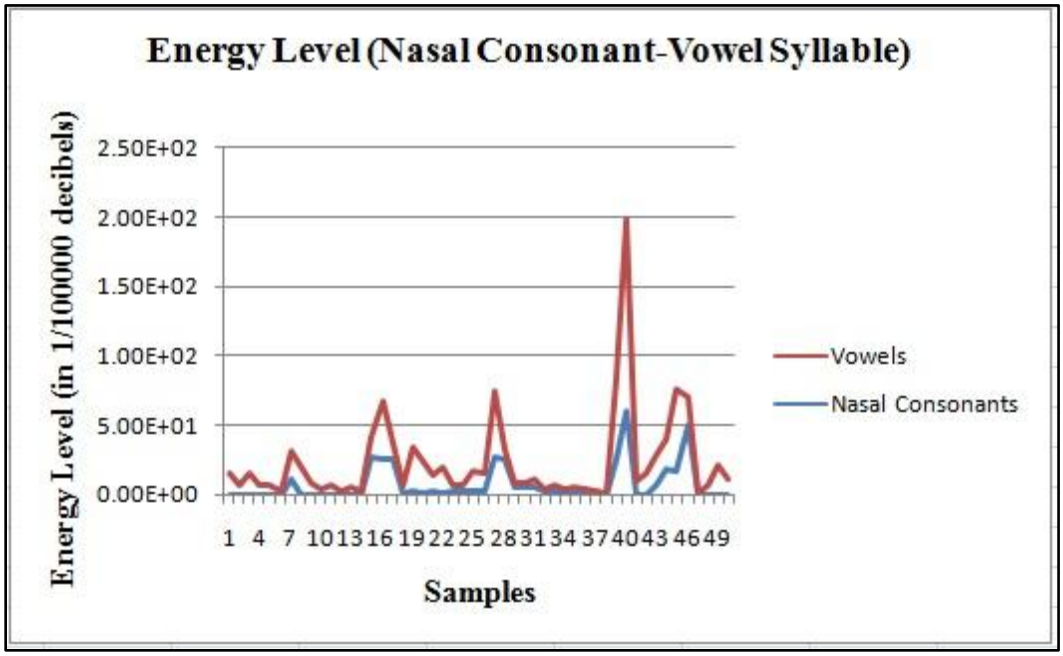

Figure 5.5: Graphs for the energy levels of Nasal Consonant-Vowel syllables

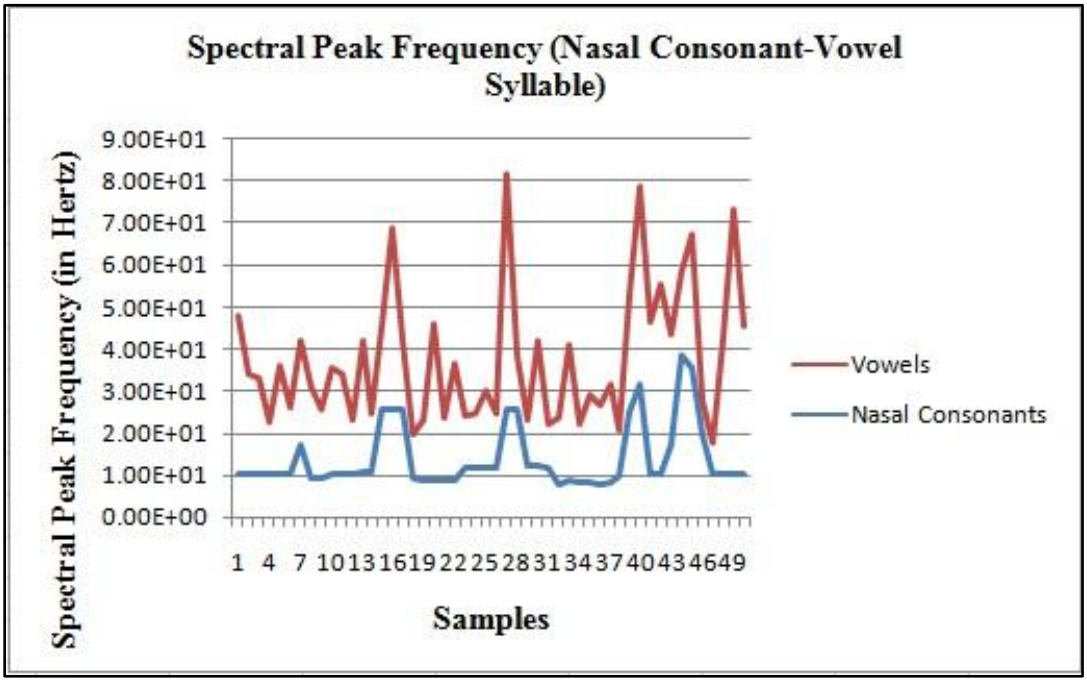

Figure 5.6: Graphs for the spectral peak frequency of Nasal Consonant-Vowel syllables

Thus, the above analysis shows that the envelope variance, energy levels and spectral peak frequency measures are lower for nasal consonants in contrast with that of the vowels.

5.2 Experimentation using LibSVM

This experiment evaluated the efficiency of two proposed techniques for automatic segmentation of syllables on a classification task. The Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables were provided as input to the Nasal Consonant-Vowel and Vowel- Nasal Consonant algorithms respectively and the nasal consonant and vowel segments of the syllable were identified by these algorithms. Then given that set of nasal consonant and vowel segments, the system had to assign a class to each of the segments. The same procedure was followed for the evaluation of the second approach to segmentation.

5.2.1 Syllable Samples collected for the experimentation

In order to carry out the classification process, Nasal Consonant-Vowel syllables were collected from 3 male speakers and 1 female speaker; and Vowel- Nasal Consonant syllables were collected from 3 male and 3 female speakers. The syllables were recorded at a sampling frequency of 16000 Hz. The Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables considered for segmentation and identification of nasal consonants and vowels are given in Table 5.1 and Table 5.2.

Table 5.1: Nasal Consonant-Vowel Syllable Samples

Nasal Consonant	Syllable (Nasal Consonant+ Vowel)
/m/	/mə/, /maː/, /me/, /mæ/, /mi/, /mɔ/, /mo/, /mu/
/n/	/nə/, /naː/, /ne/, /næ/, /ni/, /nu/, /no/, /nɔ/

Table 5.2: Vowel-Nasal Consonant Syllable Samples

Nasal Consonant	Syllable (Vowel + Nasal Consonant)
/m/	/əm/, /aːm/, /em/, /im/, /æm/, /om/, /um/, /ɔm/
/n/	/ən/, /aːn/, /en/, /æn/,/in/,/un/, /on/, /ɔn/

5.2.2 Method

Both the techniques of automatically segmenting the Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables have been evaluated on data collected in Punjabi language. The segmentation has been implemented in MATLAB using the speech processing toolbox, Voicebox. After the segmentation of Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables, the acoustic features of both the segmented portions have been used to train two SVMs (one each for Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables). The experiments have been carried out using LibSVM which is an efficient software used for SVM classification and regression. The LibSVM format of training and testing data file is:

<label> <index1>:<value1> <index2>:<value2> ...

.

.

.

Each line contains an instance and is ended by a '\n' character. For classification, <label> is an integer indicating the class label (multi-class is supported). The pair <index>:<value> gives a feature (attribute) value: <index> is an integer starting from 1 and <value> is a real number. The only exception is the pre-computed kernel, where <index> starts from 0. In the present study, label 1 has been used for nasal consonants and label 2 has been used for vowels.

5.2.3 Results

The recognition accuracy for the two approaches to automatic segmentation of syllables (Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables) has been presented below.

Approach I

In order to evaluate the technique that uses energy differences to perform segmentation of the syllables, the number of data samples for Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables collected for training and testing the SVMs and the corresponding accuracy rates have been presented in Table 5.3 and Table 5.4.

Table 5.3: Classification Experiment Result for Nasal Consonant-Vowel syllable (Approach I)

Training Data	Testing Data	Accuracy
150 syllable samples spoken by 3 male speakers	50 syllable samples spoken by 1 female speaker	65.0%

Table 5.4: Classification Experiment Result for Vowel-Nasal Consonant syllable (Approach I)

Training Data	Testing Data	Accuracy
150 syllable samples spoken by 2 male speakers and 2 female speakers	50 syllable samples spoken by 1 female speaker and 1 male speaker	91.2%

Approach II

The implementation of the first approach, that is, using energy differences to carry out the automatic segmentation of syllables provided us with just satisfactory accuracy rates for the syllables of the Nasal-Vowel category. As such another technique, which uses differences in the standard deviation of the Hilbert envelope, has been implemented on the same data set that was considered for the first technique. Table 5.5 and Table 5.6 show the number of data samples for Nasal-Vowel and Vowel-Nasal syllables collected for training and testing the support vector machines (SVMs) and the corresponding recognition accuracy in both the cases.

Table 5.5: Classification Experiment Result for Nasal Consonant-Vowel syllable (Approach II)

Training Data	Testing Data	Accuracy
150 syllable samples spoken by 3 male speakers	50 syllable samples spoken by 1 female speaker	91.0%

Table 5.6: Classification Experiment Result for Vowel-Nasal Consonant syllable (Approach II)

Training Data	Testing Data	Accuracy
150 syllable samples spoken by 2 male speakers and 2 female speakers	50 syllable samples spoken by 1 female speaker and 1 male speaker	94.0%

5.3 Comparison of the two approaches

We have discussed two different approaches to carry out the segmentation of a syllable. The temporal approach employs the differences in the energy levels of the nasal consonants and the vowels where as the spectral approach uses the difference in envelope variance of the two modes of speech. It is evident from the above discussion and the outputs of the SVM classifiers that segmentation of a syllable consisting of a nasal consonant and a vowel (for both Nasal Consonant-Vowel and Vowel-Nasal Consonant syllables) is better carried out using the differences in the standard deviation value of the Hilbert envelope of the speech signal as compared to the technique that uses the energy differences to perform the segmentation.

Chapter Summary

In this chapter, the results of the experiments that were carried out for the automatic segmentation of the Nasal Consonant-Vowel and the Vowel-Nasal Consonant syllables have been documented. The characteristics of the acoustical features: envelope variance, energy level and spectral peak frequency have been depicted graphically for the two syllable forms focused upon in this work. Further, accuracy rates of the experiments undertaken using LibSVM, for the two approaches, have been presented. This is evident from the results that the automatic segmentation of the (Nasal Consonant-Vowel and the Vowel-Nasal Consonant) syllables gives better results when carried out using the envelope variance differences in contrast with the approach that uses energy differences to perform the segmentation.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

The present work is an attempt to carry out automatic segmentation of syllables using acoustic-phonetic approach and further applying statistical learning methods in order to identify the segmented parts of the syllable. Two approaches to automatic segmentation have been presented and algorithms have been developed in MATLAB using the speech processing toolbox, namely, Voicebox. These algorithms have been implemented for the syllables (Nasal Consonant-Vowel and Vowel-Nasal Consonant) of Punjabi language. The acoustic features, namely, envelope variance, energy level and spectral peak frequency have been used to train SVM based classifier using LibSVM that recognizes the nasal consonant part and the vowel part of the syllable. Using the first approach, that is, using energy differences for the segmentation, we have been able to achieve an accuracy of 65.0% for the Nasal Consonant-Vowel syllable and 91.2% for the Vowel-Nasal Consonant syllable. Then applying the second approach, that is, using differences in envelope variance, we have been able to achieve better accuracy rates of 91.0% for the Nasal Consonant-Vowel syllable and 94.0% for the Vowel-Nasal Consonant syllable. Thus, it is evident that application of envelope variance difference facilitates better segmentation of the Nasal Consonant-Vowel and the Vowel-Nasal Consonant syllables.

6.2 Limitations of the proposed work

It is evident from the analysis documented in earlier chapters that the nasal consonants tend to have lower values for the acoustic features (envelope variance, energy level and spectral peak frequency) in contrast with the values of the vowels. Yet we have come across some cases wherein the above statement does not hold true. That is, there are some syllables wherein the nasal consonants have higher or equal values for the acoustic features than the values for vowels. This has been observed in some cases of close front vowel /i/, close back vowel /u/ and close-mid back vowel /o/. This can be noticed in Figure 6.1 of the Nasal Consonant-Vowel syllable ਮੁ /mu/ spoken by a female speaker. One can note that there is no significant onset when the vowel

articulation begins. Similar is the case in Figure 6.2 of the Vowel- Nasal Consonant syllable ਉਮ /um/ spoken by a female speaker.

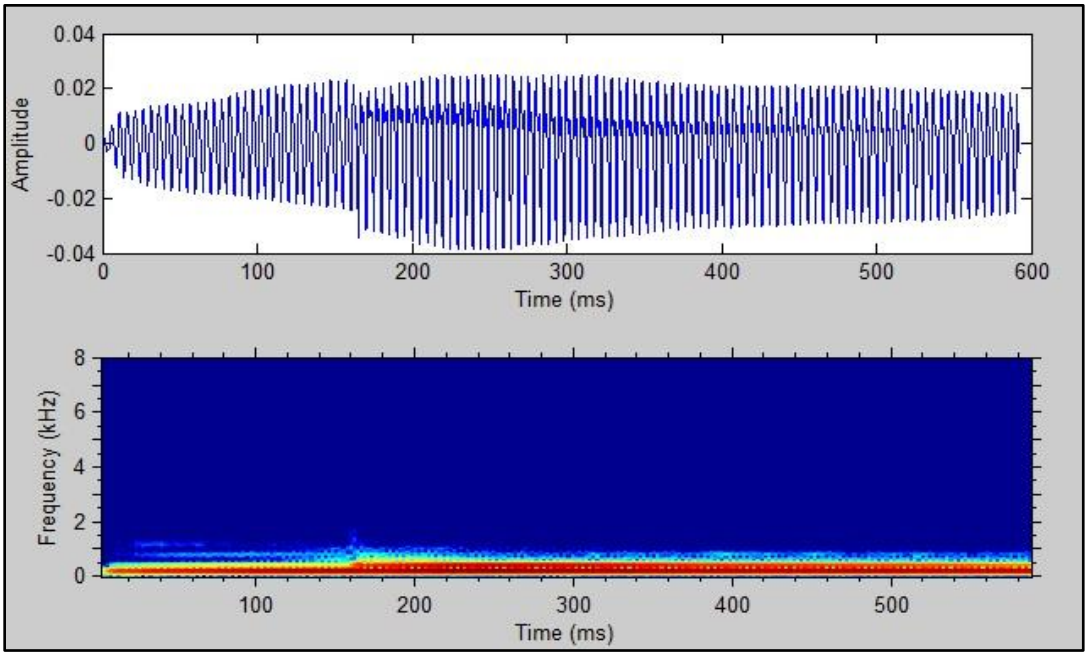

Figure 6.1: Nasal Consonant-Vowel syllable ਮੁ /mu/ a) Signal Waveform b) Spectrogram of the signal

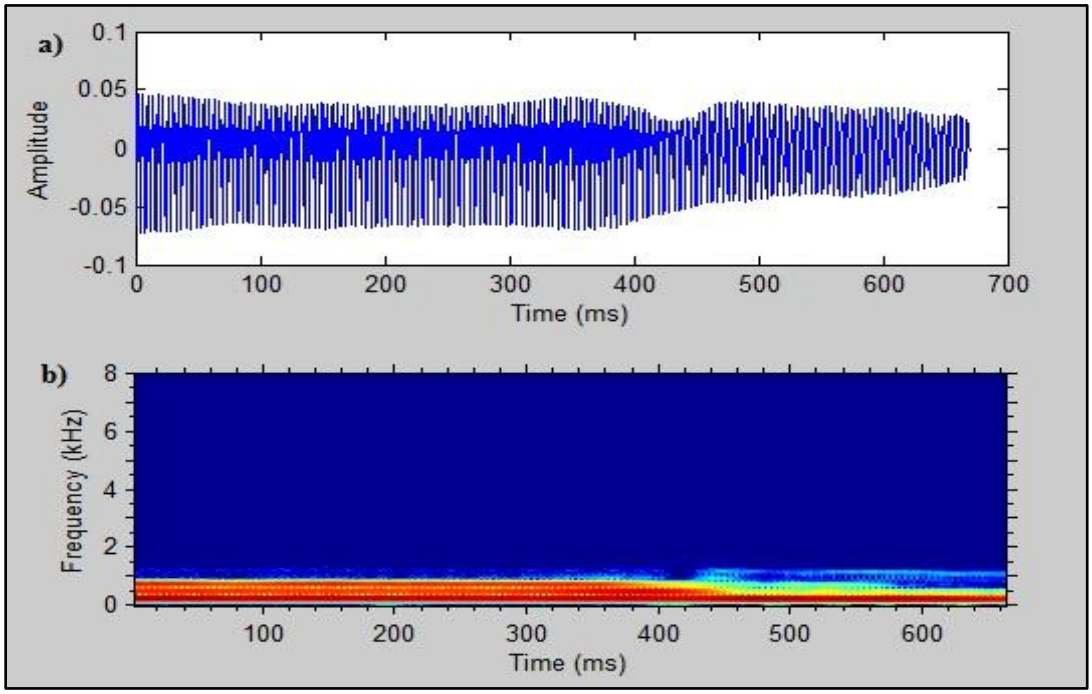

Figure 6.2: Vowel-Nasal Consonant syllable ਉਮ /um/ a) Signal Waveform b) Spectrogram of the signal

A similar absence of significant onset and offset has also been perceived in case of the syllables comprising of the vowels /i/ and /o/.

6.3 Future Scope

The focus of the present study has been on the nasal consonants (ਮ /m/) and (ਨ /n/) and the vowels (ਅ /ə/, ਆ /a/, ਐ /æ/, ਏ /e/, ਈ /i/, ਓ /o/, ਔ /ɔ/ or ਊ /u/) of Punjabi language. All of the nasal consonants and vowels have not been considered and as such, in future the nasal consonant ਙ /ŋ/ and the remaining vowels can also be examined. Also, this study has been carried out for two types of syllables: CV (Consonant-Vowel) and VC (Vowel-Consonant). Thus, it would be interesting to observe whether the present acoustic features can be applied to CVC (Consonant-Vowel-Consonant) and VCV (Vowel-Consonant-Vowel) syllables. In that case, two segmentation points would be retrieved due to the presence of an energy onset point as well as an energy offset point within the same syllable.

Some of drawbacks of the current automatic segmentation technique have been described in the previous section. In future, there is a scope for the improvement of the classification rates in case of the closed vowels. In the present work, the syllable samples, consisting of the nasal consonants and the vowels, have been recorded in a noise-free environment. Since the nasals are often confused in the presence of background noise, it would be interesting to observe how the current technique responds in case the syllables are recorded in noisy conditions.

Chapter Summary

In this chapter, the present work has been concluded. The acoustic features, namely, envelope variance, energy level and spectral peak frequency have been used to analyze the nasal consonants and the vowels at a syllabic level. The syllables of the form Nasal Consonant-Vowel and Vowel-Nasal Consonant have been focused upon for automatic segmentation.

Some of the limitations of the proposed work have also been discussed in this chapter. In general, it has been observed that the acoustic features: envelope variance, energy level and spectral peak frequency tend to have lower values for the nasal consonants as compared to the vowels. But some syllables comprising of the closed vowels such as /u/, /i/ and /o/ fail to justify the above proposition.

In future, the limitations can be overcome for the closed vowels. Moreover, the present work deals with two types of syllables: Nasal Consonant-Vowel and Vowel-Nasal Consonant. Other forms of syllables: Nasal Consonant-Vowel-Nasal Consonant and Vowel-Nasal Consonant-Vowel can also be worked upon in future. This study has been evaluated for some of the nasal consonants and vowels of Punjabi language. The nasal consonant ਘ /ŋ/ still needs to be analyzed.

References

1. Alwan A., Lo J., and Zhu Q. 1999 Human and machine recognition of nasal consonants in noise. *Proc. 14th International Congress of Phonetic Sciences,* 1: 167-170.
2. Arun V. B. 1997 A comparative phonology of Hindi and Panjabi. *Publication Bureau,* Punjabi University, Patiala.
3. Bartlett S., Kondrak G., and Cherry C. 2009 On the Syllabification of Phonemes. *Proc. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL,* Boulder, Colorado, 308-316.
4. Bitar N. N. 1998 Acoustic analysis and modeling of speech based on phonetic features. Ph.D. Thesis, Boston University, College of Engineering, Boston.
5. Bitar N. N. and Espy-Wilson C. Y. 1997 The design of acoustic parameters for speaker-independent speech recognition. *Proc. Eurospeech*, Boston, 1239–1242.
6. Brookes, M. *Voicebox: Speech Processing Toolbox for Matlab* [on line]. Imperial College, London, Available on the World Wide Web: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.
7. Brugnara F., Falavigna D. and Omologo M. 1993 Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication,* 12(4): 357-370.
8. Fujimara O. 1962 Analysis of nasal Consonants. *J. of the Acoustical Society of America*, 34(12): 1865-1875.
9. Hahn S. L. 1996 The Transforms and Applications Handbook. A. Poularakis Ed., *CRC Press,* Boca Raton, Florida.
10. Handbook of the International Phonetic Association, 1999 *Cambridge University Press*.
11. Heinzel G., Rudiger A. and Schilling R. 2002 Spectrum and spectral density estimation by the discrete Fourier transform (DFT), including a comprehensive list of window functions and some new at-top windows. Max-Planck-Institut fur Gravitationsphysik (Albert-Einstein-Institut) Teilinstitut Hannover, Germany.
12. Holmes J. and Holmes W. 2001 Speech Synthesis and Recognition. 2th ed., *Tailor Francis,* London.

13. Huang X., Acero A. and Hon H. 2001 Spoken Language Processing - A Guide to Theory, Algorithm, and System Development. *Prentice Hall PTR*, New Jersey.
14. Jabloun F., Cetin A. E. and Erzin E. 1999 Teager energy based feature parameters for speech recognition in car noise. *IEEE Signal Process. Lett.*, 6(10): 259-261.
15. Juneja A. and Espy-Wilson C. 2002 Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning. *Proc. 9th International Conference on Meural Information Processing*, Singapore, 2: 726-730.
16. Kaiser J. F. 1990 On a simple algorithm to calculate the 'energy' of a Signal. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Albuquerque, NM, 381–384.
17. Kurowski K. and Blumstein S. E. 1987 Acoustic properties for place of articulation in nasal consonants. *J. of Acoustical Society of America*.
18. Pruthi T. and Espy-Wilson C. Y. 2004 Acoustic parameters for automatic detection of nasal manner. *Speech Communication* 43: 225–239.
19. Rabiner, L. and Juang, B. H. 1993 Fundamental of Speech Recognition. *PTR Prentice-Hall*, New Jersey.
20. Ramanarayanan V., Byrd D., Goldstein L. and Narayanan S. 2010 A joint acoustic-articulatory study of nasal spectral reduction in read versus spontaneous speaking styles. *Proc. Speech Prosody*.
21. Salomon A., Espy-Wilson C. Y. and Deshmukh O. 2004 Detection of speech landmarks: Use of temporal information. *J. of Acoustical Society of America*, 1296–1305.
22. Sunil Kumar R. K. and Lajish V. L. 2012 Vowel phoneme recognition based on average energy information in the zero crossing intervals and its distribution using Ann. *International Journal of Information Sciences and Techniques (IJIST)*, 2(6).
23. Van Hemert J. P. 1991 Automatic Segmentation of Speech. *IEEE Trans. Signal Process.*, 39(4): 1008-1012.
24. Vapnik V. N. 1995 Nature of Statistical Learning Theory. *Springer-Verlag.*
25. Vorstermans A., Martens J. P. and Van Coile B. 1996 Automatic segmentation and labelling of multi-lingual speech data. *Speech Communication*, 19(4): 271-293.
26. Vuppala A. K., K. Sreenivasa Rao K. S. and Chakrabarti S. 2012 Spotting and recognition of consonant-vowel units from continuous speech using accurate detection of vowel onset points. *Circuits, Systems and Signal Processing*, 31(4): 1459-1474.