

# **SIGNAL PROCESSING FOR RECOGNITION OF ISOLATED UTTERANCES OF SPEECH UNITS**

A thesis submitted  
for the award of the degree of

**DOCTOR OF PHILOSOPHY**  
in  
ELECTRICAL ENGINEERING

by

**R . SUNDAR**

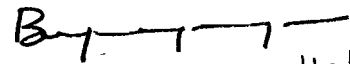


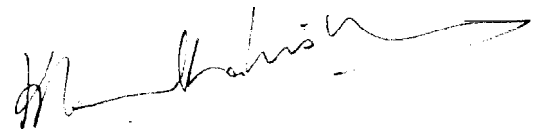
*Department of Electrical Engineering  
Indian Institute of Technology  
Madras - 600 036  
INDIA*

*JANUARY 1994*

## CERTIFICATE

This is to certify that the thesis entitled "**SIGNAL PROCESSING FOR RECOGNITION OF ISOLATED UTTERANCES OF SPEECH UNITS**" submitted by **R. SUNDAR** to the Indian Institute of Technology, Madras for the award of degree of Doctor of Philosophy is a bonafide record of research work carried out by him under our supervision. The contents of this thesis have not been submitted to any other Institute or University for the award of any degree or diploma.

  
B. Yegnanarayana 11.1.94

  
K. Radhakrishna Rao

Madras - 600 036

January 1994

## ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my guide Prof. B. Yegnanarayana. He has been a constant source of inspiration and encouragement throughout this work. I am indeed fortunate to have him as guide and advisor. But for his interest and motivation, it is unlikely that this thesis could have taken shape. I have benefited immensely through his association. I sincerely thank him for all the help he has rendered.

I wish to express my gratitude to my co-guide Prof. K. Radhakrishna Rao and I sincerely thank him for his support and encouragement throughout. I wish to thank Dr. S. Raman, my friend and advisor who initiated me to this research work.

I express my sincere thanks to Prof. (Mrs) Kamala Krithivasan, Head of the Department of Computer Science and Engineering for providing the facilities to carry out my research work. I wish to thank Prof. H.N. Mahabala and Prof. R. Nagarajan for their interest, encouragement and help. I express my sincere thanks to all my colleagues in the Department of Computer Science and Engineering.

Special thanks are due to my friends Mr. C.ChandraSekhar and Dr.P.Eswar who have made numerous constructive criticism in improving the presentation of the thesis. I am very grateful to them for the help and assistance they have rendered during the preparation of the thesis. I can never repay their timely help. I thank Dr. (Mrs) Hema A. Murthy and Mr. G.V. Ramana Rao for their help and useful comments.

I am bereft of words to express my thanks to Mr. Alwar who made my worries less by taking the burden of preparing the final dump and copies. I have taken his help for granted and I am indeed very grateful for his timely help.

I thank Mr.S.Rajendran Mr. Ramaseshan who went out of their way to help me. My thanks to Hashim, Tamilendran for their help in the preparation of tables and figures.

I thank my friends in Speech and Vision lab and Microprocessor Lab for the help they rendered.

Finally, I wish to thank my FAMILY and FRIENDS for their support and help in this endeavor.

*R. SUNDAR*

# CONTENTS

## ABSTRACT

### 1. INTRODUCTION TO SIGNAL PROCESSING ISSUES

<b>IN SPEECH RECOGNITION .....</b>	<b>1</b>
1.1 Background to recognition of isolated utterances .....	2
1.2 Scope of the present study .....	4
1.3 Issues addressed and studies carried out .....	6
1.4 Organization of the thesis .....	11

### 2. REVIEW OF METHODS FOR ISOLATED WORD

<b>SPEECH RECOGNITION .....</b>	<b>13</b>
2.1 Isolated word recognition strategies .....	13
2.2 Knowledge-based signal processing for speech recognition	21

### 3. ACOUSTIC-PHONETIC KNOWLEDGE OF CV UTTERANCES

<b>IN HINDI .....</b>	<b>24</b>
3.1 Introduction to acoustic-phonetics .....	24
3.2 The Hindi alphabet .....	25
3.3 Speech production mechanism .....	27
3.4 Nature of speech signal in CV utterances .....	30
3.5 Acoustic features of CV utterances .....	32
3.5.1 Features of broad classes and regions in CV utterances .....	32
3.5.2 Acoustic features of regions .....	36
3.5.2.1 Voicing .....	36
3.5.2.2 Stop release transient .....	37
3.5.2.3 Aspiration .....	38

3.5.2.4	Voiced aspiration .....	39
3.5.2.5	Nasal .....	39
3.5.2.6	Semivowel .....	40
3.5.2.7	Frication .....	40
3.5.2.8	Vowel-like .....	41
3.6	Methodology for recognition based on acoustic-phonetics	41

#### **4. DETECTION OF SIGNIFICANT INSTANTS FOR**

<b>IDENTIFICATION OF DIFFERENT REGIONS .....</b>	<b>48</b>
4.1 Significant instants in CV utterances .....	49
4.2 Issues in determining significant instants .....	50
4.2.1 Start of the utterance .....	52
4.2.2 The articulatory release instant .....	52
4.2.3 The vowel onset instant .....	53
4.2.4 Summary .....	53
4.3 Epoch detection to identify significant instants .....	54
4.4 Determination of instants based on parametric change ....	58
4.5 Discussion of results .....	59

#### **5. STUDIES ON RECOGNITION OF VOWELS IN CV UTTERANCES 62**

5.1 Issues in vowel discrimination .....	63
5.1.1 Manifestation of speaker characteristics .....	63
5.1.2 Cues for vowel discrimination .....	64
5.2 Choice of parameters for vowel classification .....	66
5.2.1 Gross spectral features for vowel classification ..	66
5.2.2 Parameters for vowel discrimination used in this study .....	70
5.3 Vowel classification using a neural network classifier .....	71
5.3.1 Data preparation for the classifier .....	74

5.3.2	Recognition studies .....	75
5.4	Discussion of results .....	75
<b>6.</b>	<b>STUDIES ON RECOGNITION OF CONSONANT MANNERS ...</b>	<b>87</b>
6.1	Distinguishing features of consonantal regions .....	88
6.2	Choice of parameters for consonant region discrimination	94
6.2.1	Waveform periodicity .....	94
6.2.2	Relative energy .....	95
6.2.3	Duration of region .....	95
6.2.4	Spectral structure .....	96
6.2.5	Relative high frequency energy .....	96
6.2.6	Gross spectral nulls .....	96
6.2.7	Summary .....	97
6.3	Classification of consonantal regions of CV utterances ....	97
6.4	Discussion of results .....	100
<b>7.</b>	<b>STUDIES ON RECOGNITION OF PLACE OF ARTICULATION</b>	<b>106</b>
7.1	Issues in analyzing the dynamic region .....	107
7.2	Choice of parameters to detect place features .....	108
7.3	Methodology for recognition .....	110
7.4	Studies on the determination of place of articulation .....	111
7.5	Discussion of results .....	112
<b>8.</b>	<b>STUDIES ON OVERALL RECOGNITION OF CV UTTERANCES</b>	<b>115</b>
8.1	Classification studies of each region using parameters derived from the region alone .....	116
8.2	Classification studies of each region using parameters of two adjacent regions .....	120
8.3	Classification studies of each region using parameters of all three regions .....	123

8.4	A hierarchical model for the recognition of CV utterances	128
8.5	Summary .....	130
9.	<b>SUMMARY AND CONCLUSIONS .....</b>	<b>131</b>
	<b>REFERENCES .....</b>	<b>136</b>
	<b>LIST OF PUBLICATIONS .....</b>	<b>145</b>

## LIST OF FIGURES

- Fig. 2.1** : An example of typical warping function
- Fig. 2.2** : Speech signal and spectrogram for a segment of utterance [ka]
- Fig. 3.1** : Sketch of the parts of human speech production mechanism
- Fig. 3.2** : Speech waveform for the utterances  $[k^h]$  and [tʃ]
- Fig. 3.3** : Significant instants and regions for the different manners of consonants
- Fig. 4.1** : Plots of speech waveforms and gross parameters for utterances of Hindi stop consonants.
- Fig. 4.2** : Speech waveform and the epochs (instants of significant excitation) derived using the phase derivative method for some CV utterances
- Fig. 5.1** : Hindi vowels in F1-F2 plane
- Fig. 5.2** : LP spectrum and different normalizations of the area functions for many repetitions of different vowels by a male speaker
- Fig. 5.3** : LP spectrum and different normalizations of the area functions for many repetitions of different vowels by a female speaker
- Fig. 6.1** : Histograms of some parameters in different regions of CV utterances



## LIST OF TABLES

- Table-3.1 : The Hindi alphabet arranged to reflect classification based on acoustic-phonetic features
- Table-3.2 : Broad classes of Hindi CV utterances based on the manner of production of consonants
- Table-3.3 : Significant regions in broad classes of utterances
- Table-3.4 : Time sequences of instants and regions in broad classes of utterances
- Table-4.1 : Parameters for characterizing significant instants
- Table-4.2 : Performance of identification of significant instants .  
in percentage values
- Table-5.1 : Performance of vowel discrimination - Speaker A
- Table-5.2 : Performance of vowel discrimination - Speaker B
- Table-5.3 : Performance of vowel discrimination - Speaker C
- Table-5.4 : Performance of speaker independent vowel discrimination
- Table-5.5 : Performance of multispeaker vowel discrimination
- Table-5.6 : Summary of correct classification performance for different parametric representations
- Table-5.7 : Summary of correct classification performance for different vowels
- Table-5.8 : Vowel classification considering the best two candidate classes -  
Speaker A
- Table-6.1 : Characteristics of features
- Table-6.2 : Feature characterization in different regions
- Table-6.3 : Distinguishing features of regions
- Table-6.4 : Parameters for region classification
- Table-6.5 : Manner classification performance

- Table-6.6 : Performance of manner classification using reduced set of parameters
- Table-6.7 ,: Performance of manner classification including trend parameters
- Table-6.8 : Performance of manner classification in the context of different vowels
- Table-7.1 : Performance of classification of place of articulation
- Table-8.1 : Summary of the classification performance for the three parts of the CV utterance
- Table-8.2 : Classification of place of articulation using all the five transition frames with ten parameters per frame
- Table-8.3 : Classification of place of articulation using three transition frames with six parameters per frame
- Table-8.4 : Manner classification using both manner region parameters and transition region parameters
- Table-8.5 : Classification of place of articulation using both transition region parameters and manner region parameters
- Table-8.6 : Classification of place of articulation using both transition region parameters and vowel region parameters
- Table-8.7 : Vowel classification using both transition region parameters and vowel region parameters
- Table-8.8 : Manner classification using all three regions' parameters
- Table-8.9 : Vowel classification using all three regions' parameters
- Table-8.10: Classification of place of articulation using all three regions' parameters
- Table-8.11: Overall classification using all three regions' parameters with a single network having all the three types of outputs

## ABSTRACT

This thesis reports our studies on the recognition of isolated utterances of Hindi characters. Hindi characters are composed of mostly stop consonants and a few other types of consonants and pure vowels. The importance of signal processing in the context of recognition of short units of speech is highlighted. Acoustic-phonetic knowledge has been used as the basis to evolve an approach for recognition. This knowledge is also used to identify features that carry distinctions among different categories of sounds. Three distinct regions in each utterance of a character are identified for detailed analysis. The region before the vowel onset is used to determine the consonant group. The region after vowel onset is used to determine the vowel type, and the transition region - the short (30 msec.) region immediately after the articulatory release - is used to determine the place of articulation of the consonant. It is shown that these three different parts need to be processed differently for recognition of the utterance of a character. In particular, the region prior to the vowel onset is mostly of transient type, and hence it is more appropriate to represent it by gross spectral and temporal features. The vowel region is to be described in terms of either the steady articulatory parameters or the resonances corresponding to the steady vocal tract system or the parameters of a linear system model for the vocal tract system. The transition region is more aptly described by a sequence of parameter values reflecting the dynamics of the vocal tract system. Methods based on characteristics of group delay functions are applied to obtain the significant instants separating these regions accurately for manual identification. Parameters which reflect the vocal tract shape namely, area coefficients functions are studied to represent the vowel part of the utterance. The objective is to explore the use of such parameters related to the vocal tract shape to take care of variability due to speaker characteristics. Area coefficients are also used to represent the transition region with the hope of characterizing it by the changes in area coefficients. The features in the transition region reflect the context of the consonant manner and

the vowel type of the utterance. Therefore, the recognition studies for the place of articulation is carried out in the context of consonant manner and the vowel class. The main contributions of this research work are in evolving a methodology for the recognition of CV utterances and in proposing specific choices of parameters for representing different regions of the utterances.

## **INTRODUCTION TO SIGNAL PROCESSING ISSUES IN SPEECH RECOGNITION**

The objective of this thesis is to demonstrate the significance of signal processing for speech recognition, especially for isolated utterances of small units of speech. We discuss this with special reference to utterances of isolated characters of the Indian language, Hindi. In the absence of clues due to redundancy of speech and language as in continuous speech, recognition of isolated utterances has to rely on the features that can be extracted from the signal and the acoustic-phonetic knowledge pertaining to the speech units of the language. We show that the use of acoustic-phonetic knowledge in signal processing, feature extraction and matching is necessary for improving the recognition accuracy significantly over the conventional approach [Rabiner 1993] of using uniform parametric representation and matching strategies based on these representations.

For the utterances of characters of Hindi, different regions in the signal were identified based on the acoustic-phonetic knowledge. The signals in these regions were processed appropriately to extract parameters and features suitable for recognition of these regions. The features include spectral energies, temporal characteristics, area functions and several others. The recognition studies were made using trained neural network models for classification. The studies show that if proper attention is not given to the signal processing stage, there is hardly any possibility of achieving useful recognition performance for such complex tasks as the recognition of isolated utterances of characters of a language.

In the following discussion we highlight the background, scope of the current work, and identification of issues for detailed study.

## **1.1 BACKGROUND TO RECOGNITION OF ISOLATED UTTERANCES**

Recognition of isolated utterances of speech units is normally performed by extracting parameters for each fixed duration segment of speech signal and matching the sequence of parameter vectors for test and reference utterances [Lea 1980] [Rabiner 1981] [Rabiner 1993] . The parameters are usually related to spectral information in the segment. The matching is performed using either dynamic time warping (DTW) or by a stochastic model approach like hidden Markov model (HMM) [Rabiner 1989]. This takes care of compression and expansion of different regions in the test and reference patterns. These methods work reasonably well for nonconfusable multisyllabic words. Their performance degrades significantly if the utterances are confusable and short, as in monosyllabic units. The main reason for their failure is that the parametric stage does not capture the significant features of production of these units. Moreover, the spectral shapes for similar segments are different for different speakers due to differences in the shapes of their vocal tract system, [Fant 1970] [Flanagan 1972].

For recognition of isolated utterances of characters of a language, the parametric representation of the speech signal becomes crucial [Raman 1985] [Sundar 1987] [Yegnanarayana 1986]. This is the reason for the poor performance of recognition of systems for English alphabet task [Yegnanarayana 1984]. The problem is further compounded for the character set in Indian languages due to the size of the character set.

The task of developing a recognition system for the utterances of isolated characters of Hindi differs from the task of developing such a system for the alphabet in English in several ways. The alphabet of English is relatively much smaller than the Hindi character set. The confusability in the alphabet set of English is confined to a subset of the vocabulary whereas the confusability is distinctly higher and occurs in many groups of the Hindi character set. In the Indian language context, recognition of isolated utterances of characters can form the basis for a character-by-character input which is equivalent to very slow speech. This is possible due to the phonetic nature of the language wherein the spoken form and the written characters have nearly a one-to-one correspondence. Due to this phonetic nature, the utterances of isolated characters in sequence, represent the message of continuous speech although there are artificial gaps. These features imply that studies carried out on isolated character recognition for the Indian languages may be useful to some extent in the study of continuous speech recognition problem as well. Another feature is the absence of effects due to coarticulation in isolated character utterance, whereas in continuous speech the coarticulation effects significantly influence the parameters of speech.

However, studies on isolated word recognition do not naturally lead to continuous speech recognition. This is because isolated word speech recognition is usually approached as a pattern matching problem, where the constraints imposed by the language as well as the constraints imposed by the speech production mechanism are not generally used.

The study of isolated character utterance recognition has importance in its own right as it can be used in many applications such as learning aid, directory assistance, keyboardless data entry and even in automated speech therapy sessions for hearing impaired persons. The Indian languages have a large character set. Normal keyboard based data entry is cumbersome as it calls for

multiple keystrokes for each character. Keyboardless data entry through speech overcomes this problem.

Isolated utterances of characters of the Hindi language are a confusable set. The parameter matching approach does not lead to a satisfactory solution. For example, the difference between the utterances of characters alveolar [t<sup>^</sup>] and denti alveolar [T<sup>^</sup>], is marginal and cannot be brought out by dynamic time warping and distance computation. One has to use information about speech production mechanism and language features to discern that these are distinct utterances.

## **1.2 SCOPE OF THE PRESENT STUDY**

The main objective of the present study is highlighting the issues in realizing a recognition system for isolated utterances of characters of Hindi. Speech mode of communication with a machine involves two major components. One is the recognition of characters from speech signal and the second component is the protocol for interaction with a machine in speech mode. Our interest is in the problem of recognition of isolated utterances of characters. From a practical point of view, the scope of the work should extend to handle any speaker's utterance in a normal environment such as an office room having a reasonably low level of background noise. An implicit assumption in this work is that isolated utterances are, in general, clearly spoken. This implies that all features of production of the character are present in the signal. In contrast, features of characters in continuous speech are modified by neighboring sounds to the extent of even being absent in some cases. The articulatory description for the production of characters (and hence the corresponding sequence of acoustic events) is generally unique for an Indian language. Thus there is very little scope for variability in the description of features of character utterances from speaker to speaker.



Isolated character utterances have less variability due to absence of context, but, variability in features due to different speakers needs to be accounted for. A disadvantage in isolated character utterance is that trends in features due to speaker characteristics, which might be evident in continuous speech context, are not available in isolated utterances.

The absence of context of neighboring speech regions in isolated character utterances implies that one cannot take advantage of the higher level knowledge sources such as lexical, syntax, semantics and prosodic which impose their own constraints in continuous speech. It is also not possible to make use of higher level phonetic knowledge arising out of coarticulation. Therefore, for recognition, it is not possible to use some approximate representation (in arbitrary symbols such as phones) of speech signal and expect that higher level knowledge sources to disambiguate the sequence of phones [Sundar 1987]. Recognition of the characters is to be performed by processing the signal to extract the acoustic features accurately.

The character set consists of vowels (**V**), consonant-vowels (**CVs**) and consonant clusters followed by a vowel (**C<sup>\*</sup>V**). A point of note is that the speech production rules for these units are unique, requiring similar manner, and places of articulation and thus these rules are speaker independent. However, their manifestation in the parameters of the speech signal is significantly dependent on the characteristics of the speaker's vocal tract system. Moreover, some of these production units are very close in terms of place of articulation as in [t<sup>^</sup>] and [T<sup>^</sup>] in Hindi, for example. Therefore, processing of speech signal without taking into account the manner and place of articulation of production is likely to give parameter values which do not reflect these distinctions. Sophistication in the subsequent matching for recognition is not likely to overcome their deficiency. The vocabulary for the present task consists of CV utterances formed **by** combining the thirty three consonants with the ten vowels of Hindi. Thus the total number

of characters is three hundred and thirty. The utterances are pronounced clearly as per the rules of production in a reasonably noise free environment, and recorded by a microphone kept close (maximum two inches away) to the speaker. Thus it is assumed that all the production features are carefully articulated and captured. The ultimate goal is to realize a speaker independent recognition of these utterances. Throughout, the signal processing and the choice of parameters are dictated by these considerations.

### **1.3 ISSUES ADDRESSED AND STUDIES CARRIED OUT**

The following issues [Yegnanarayana 1991] were identified for study in this thesis:

- 1) Since the CV units are dynamic sounds, it is necessary to identify the instants (regions) where major acoustic events take place. For example, in the case of unvoiced velar stop [k], the instant of burst release, the onset of voicing and the transition region are the important events [Harrington 1988]. An automatic method of detecting these instants/regions is the first main issue to be addressed. The three important regions for CV utterances are the region before vowel onset instant, the region during transition from vowel onset to steady vowel and the steady vowel region. There are a few exceptions to this, as in the aspirated sounds, which would be considered at the appropriate stage of analysis and recognition.
- 2) The vowel type is to be determined by using the features representing the vocal tract shape in the vowel region. These features are derived through analysis of the spectral characteristics in the vowel region.

- 3) Each of the identified regions of the speech signal should be processed and parameters/features relevant to that region have to be extracted for recognition. The region before vowel onset is used to determine the manner class of the consonant such as unvoiced(voiced)–unaspirated, unvoiced(voiced)–aspirated, nasal, fricative and semivowel.
- 4) The transition region corresponding to a short (about 30msec) duration after vowel onset is to be analyzed for recognition of the place of articulation. The analysis requires extraction of the dynamic characteristics of the vocal tract shape in the region.
- 5) While each of the above issues can be addressed separately to some extent, there is significant influence of the features of one region on the other. Hence, the above recognition studies result in multiple hypotheses for manner, vowel type and place of articulation. These need to be resolved by combining the evidence from all the regions together to enhance the overall recognition performance.

In this thesis each of these issues is addressed separately, using manual segmentation of the speech signals and the acoustic phonetic knowledge of the other regions, wherever needed, to study the identified issue more thoroughly. This is done primarily to bring out the importance of appropriate signal processing in each region for the purpose of recognition. Recognition studies are made using neural network models for classification [Lippman 1987] [Hush 1993]. Though signal processing is done considering the features that need to be detected, the features are usually buried in the variability of the parameters. Using the neural networks approach, it is hoped that the mappings based on these features can be learnt from examples.

Different approaches have been adopted for the identification of boundaries of different regions in speech signal [Rabiner 1978b] [Rabiner 1993]. In our approach, a trained neural network was used to identify boundaries of different regions of CV utterances. The network was trained using a set of appropriate parameters for either side of the boundary. In fact, these boundaries correspond to significant instants of excitation of source. A method based on characteristics of phase derivative or group delay function has been applied to determine accurately the instants where significant changes take place in production [Smits 1992]. The instants of significant excitation identified for some CV utterances using this method were used as an aid for manual segmentation of regions prior to studies on individual regions. These were also used to prepare training data for a neural network classifier.

For recognition of the vowel regions -it is necessary to use parameters/features that reflect the shape of the vocal tract system. Since this is a large amplitude region, it is possible to reliably extract the spectral features. However parameters related to vocal tract shape such as area coefficients are studied in detail. The area coefficients are extracted by deriving linear predictor coefficients (LPCs) from speech signal [Wakita 1973]. LPCs can be reliably extracted from speech data in these regions due to the availability of high signal to noise ratio (SNR) signal. We have obtained a good recognition accuracy for vowels using a neural network classifier, confirming the suitability of the area coefficients, although other parameters such as cepstral coefficients seem to offer better accuracies. In all the cases even when there are errors in recognition, the shape of the vocal tract for the confused vowel is close to the shape of the vocal tract for the true vowel. Thus, the neural network classifier can be used to obtain the best and next best choices; in case of close decisions, for later-processing.

For determining the manner of production from the region before vowel onset, the following points were considered. The region primarily consists of

noiselike signal due to sound production from narrow constriction in the vocal tract system, except in certain cases such as voiced region in voiced CV utterances. Therefore, parameters relating mostly to temporal and gross spectral features only were considered. Since this region typically has relatively small amplitude as compared to the signal in the following vowel region, the spectral parameters, even if they are important in some cases, cannot be relied upon due to sensitivity of the spectral parameters to the inevitable additional noise in these regions. We have shown that using parameters based on energy, zero crossings count, spectral flatness, duration and such other gross parameters, a trained neural network gives good recognition performance. Even when there were errors, the recognized class was close to the correct one in terms of features. As an example, the broad consonantal class voiced-aspiration is mostly confused with the aspiration class. Similarly, the semivowel class is confused with the nasal class. In both of the above cases, the confused classes show distinct commonality of features. Further improvement is possible only if a better set of parametric representation can be found. In all these studies, manually segmented regions were used.

Recognition studies for the place of articulation were made using the transition region corresponding to a fixed (30 msec.) duration after vowel onset. For aspirated sounds, the transition region was considered just after the release of aspiration. Since the parameters have to capture the dynamic characteristics of the vocal tract system, the area coefficients for several consecutive frames in this region were tried as input to a neural network classifier. Since the dynamics of the vocal tract system depends not only on the place of articulation, but also on the following vowel, it was necessary to take into account all the cases of vowels for each place of articulation while designing the neural network classifier. For this purpose we have developed a multiclass neural network classifier. We have noticed that representation of each frame of data by cepstral coefficients

derived from LPCs yielded better results for place recognition than by the area coefficients which is largely due to variability in the computed area coefficients.

Finally, an overall recognition strategy is proposed based on the output from the studies in each region. Since recognition studies of each region produce multiple hypotheses, it is possible to check the consistency of these hypotheses across all regions in the utterance to arrive at a final decision on the overall recognition of a CV utterance.

The specific contributions of this research work are summarized below.

1. The major contribution is in proposing specific choices of parameters for representing different regions for classification purposes. In particular,
  - a) Only gross temporal and spectral features are needed for classification of manners of production of consonants.
  - b) The cepstral or area coefficients derived from LPCs are found to be useful for achieving a speaker dependent recognition of vowels.
  - c) The cepstral or area coefficients are also useful for identification of place of articulation from the transition region.
2. Methods based on phase derivative properties and neural network models were proposed for identification of significant instants and regions in the speech signal for CV utterances.
3. A hierarchical model is proposed for recognition of any of the 140 CV consonants from its utterance. The model gives an estimated overall CV recognition of 65% which is significantly high considering the fact that we have chosen a large (140) set of highly confusable CV utterances.

## 1.4 ORGANIZATION OF THE THESIS

In Chapter 1, we have dealt with the background to the problem and scope of the present work. The state of art in isolated word recognition and the need for knowledge driven approach to this recognition problem particularly concerning the confusable vocabulary of CV utterances of the language Hindi are given in Chapter 2. In Chapter 3, we discuss the acoustic phonetic knowledge of Hindi which provides the link between speech production mechanism and the resulting signal. Important acoustic features and their manifestation in the different categories of sounds are also discussed. This chapter also details the categorization of CV utterances which paves the way for a suitable approach to realize recognition. In this chapter, we also discuss methods to achieve high recognition accuracy by using features as derived from acoustic-phonetics. This chapter highlights the importance of addressing the issue of feature detection and consequently that of signal processing. The end of this chapter (chapter 3) identifies the four separate problems that need to be addressed to recognize the CV utterance. Each of the four problems is taken up as a separate study in chapters 4, 5, 6 and 7. In chapter 4, the methodology for determining the significant instants in the CV utterance is evolved and discussed. Use of special signal processing methods to determine these instants is also discussed. In the last part of chapter 4, detection of significant instants using gross parameters as input to a neural network classifier is discussed. Chapter 5 addresses the issues in determining the vowel type of the CV utterance. The signal from the region after vowel onset in the CV utterance is used as the relevant region for analysis. The nature of signal processing and the choice of parametric representation needed for characterizing the vowel type is discussed. At the end of this chapter, the vowel type determination using the chosen parametric in a neural network classifier is discussed. In chapter 6, the determination of the manner of the consonant of a CV utterance is studied. The region in the CV utterance before vowel onset is identified as the relevant region for analysis. Chapter 7 addresses

the issues in analysing the dynamic region corresponding to the transitions arising out of the consonantal release. The place of articulation of the consonant is determined by analysing the region immediately following the articulatory release. Chapter **8** addresses the recognition of overall CV utterance using the result of the studies carried out in chapters **4**, **5**, **6** and **7**. Chapter **9** summarizes the work and highlights the contributions. It also discusses some issues for further study.



## **REVIEW OF METHODS FOR ISOLATED WORD SPEECH RECOGNITION**

Isolated word speech recognition systems normally work for a limited vocabulary of words. If words in the vocabulary are multisyllabic, the gross dissimilarities in the syllable sequence of different words are adequate to disambiguate the words. Such a vocabulary can be termed nonconfusable. Performance of **IWSR** systems is mainly dependent on the size and nature of the vocabulary. This chapter discusses different methods used for **IWSR**, highlighting the basic concepts of pattern matching techniques as applied to this problem. Drawbacks of direct pattern matching approach and approaches based on hidden Markov model and neural network models are discussed. Deficiencies in the conventional signal processing methods used for **IWSR** are also brought out. The need for signal-dependent as well as knowledge-based approach are discussed leading to a discussion on the need for acoustic-phonetic knowledge in signal processing, particularly in the context of CV utterances.

### **2.1 ISOLATED WORD RECOGNITION STRATEGIES**

Isolated word recognition strategies generally incorporate the pattern matching technique [Rabiner 1981] which exploits the gross dissimilarities among words. The features of a word are obtained by parametric representation of speech signal. The time domain speech signal is divided into time slots called frames. The spectral characteristics of the signal in these frames are obtained as parameters. The time sequence of these parameters reflects the variations that

occur in the speech signal of a word. During the training phase, these features for typical utterances of words in the vocabulary are obtained and stored as reference patterns. During the recognition process, similar features obtained for a test word are matched against these references. The matching strategy is based on some form of distance measure between the test word and the reference word. The distance measure is computed by taking the integrated difference between suitably weighted parameter values of the test word and the reference word over appropriate frames.

Choice of parameters is an important consideration as it determines how well the relevant variations in features are captured. The best parameter choice among a set of parametric representations for optimal performance of speech recognition systems has been studied and reported in [Davis 1980].

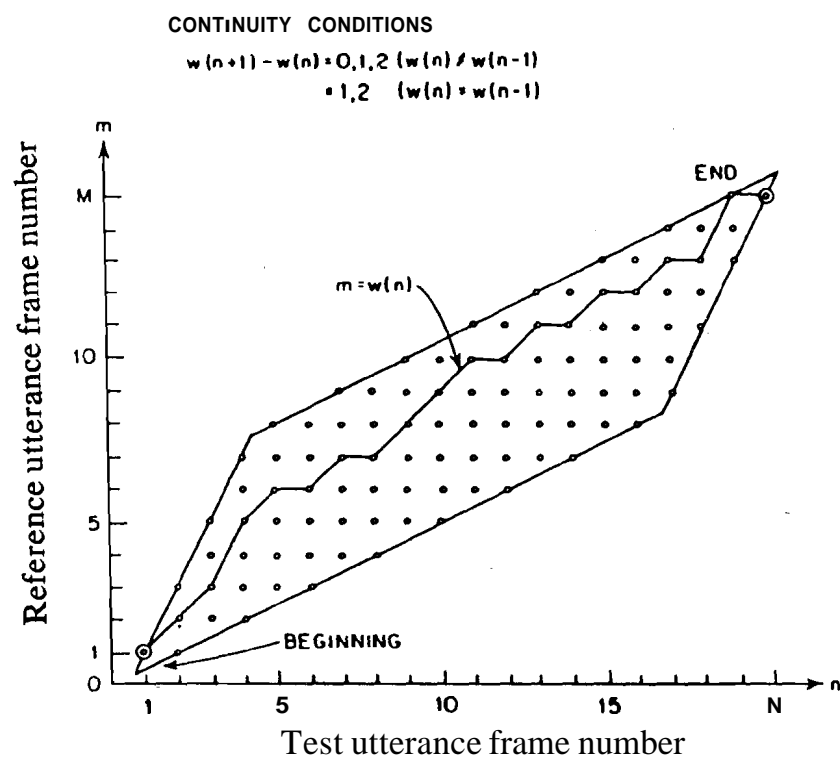
Conventional approaches use what may be termed as fixed strategies in parametrization, namely, uniform frame size and fixed set of parameters. This is because, mechanism for incorporating context dependent parameters makes the process of comparison very complex and would lack common metric to measure the closeness.

Speech is characterized by inherent variability. An important and conspicuous variability is the one that arises out of variations in the rate of speech. This variation results in changes in the duration (time compression and expansion) of acoustic features in different utterances of a word uttered at different instances by the same speaker. To overcome the errors in recognition due to this variation, time normalization is carried out as part of distance measure computation when comparing the test word features against the stored reference word features. This time normalization or rather time alignment between the test and reference features is carried out by the use of a dynamic programming algorithm, more commonly called Dynamic Time Warping (DTW) algorithm. The DTW algorithm attempts to

try out various paths of time compression and expansion and obtains the consequent distance scores. The path that yields the least distance score is presumed to be the best time normalization between the test and reference words. Since trying out all possible paths leads to a time consuming search, and, since physical constraints of the speech production mechanism limits the extent of durational variations, DTW algorithm implementation normally includes bounds on the extent of allowed time dilation and contraction. The limits are applied both in local (at any instant within word) as well as global (overall duration of word) context. Fig. 2.1 shows a typical warping path between a test word and a reference word. The figure also shows the typical constraints applied when performing the DTW. Several variations are possible in the implementation of DTW algorithm. The optimal choice of a DTW algorithm has been extensively studied and reported by [Sakoe 1978] [Myers 1981].

In the context of speaker independent recognition, variability due to speaker characteristics is an important issue. To obtain good recognition performance under this situation either a training approach is used where the references for a specific speaker are obtained first before comparison, or multiple references are obtained from different speakers and they are used judiciously [Rabiner 1978a].

Another approach to handling the durational variations in IWSR is by the use of hidden Markov models (HMM) [Rabiner 1989]. In this approach, each word in the vocabulary is modeled by a distinct HMM. The HMM for each word is obtained by estimating the model parameters using training patterns. The gross time sequence of events in the word are captured in the hidden state sequence of the model. The durational variations are absorbed in the state transition probabilities of the model.



**Fig. 2.1 : An example of typical warping function**

Recently, there have been many attempts to use artificial neural networks for speech recognition tasks. The artificial neural networks are interesting for cognitive tasks such as speech and vision that are characterized by a high degree of uncertainty and variability. They offer ways of automatically designing systems that can make use of multiple interacting constraints which are too complex to be explicitly stated. The capabilities of multilayer perceptron models to learn automatically from examples, to form complex decision surfaces and to generalize from what is learnt, make these models suitable for classification tasks [Lippmann 1987]. Multilayer perceptrons that are trained in supervised mode using backpropagation algorithm have been used for isolated word recognition [Burr 1988]. These models have very good discrimination capabilities but their use for classification of patterns of varying duration, such as in speech, is limited. This limitation for using these models for isolated word recognition is overcome by performing nonlinear time alignment using trace segmentation [Demichelis 1989] or by compressing the parametric representations of the words to obtain fixed duration patterns [Freisleben 1992]. A translation-invariant model based on time-delay neural network architecture was used for recognition of unaligned patterns [Lang 1992]. Applying the prior knowledge of the task and its properties is important for successfully developing a neural network classifier for tasks such as speech recognition.

The isolated word recognition systems with limited vocabulary are used for a wide variety of real world applications. Attempts are also being made to develop systems for recognition of spoken letters or characters of a language [Cole 1990] [Lang 1992]. These systems can be used for applications like directory name retrieval in English [Cole 1991] or for input of characters into computers when the other modes of input, such as keyboard are difficult to use. The latter application is relevant for the languages, such as Chinese and Indian languages, which have a large number (in the range of 5000 to 10000) of characters. The work on development of a speech dictation machine for Chinese language with

large vocabulary is reported in [Lee 1993]. In the present thesis work, we address some issues in the recognition of isolated utterances of characters in Hindi, an Indian language.

The isolated spoken letter or character recognition systems are characterized by their large and highly confusable vocabularies. Therefore, the conventional approaches used for isolated word recognition of limited and nonconfusable vocabularies cannot be directly adopted. The approaches for developing isolated letter or character recognition systems with good classification performance, should use parametric representations which can capture the characteristics of different segments of the isolated utterances, and use classifiers with good discrimination capabilities.

The choice of simple parameter matching works for the general class of recognition problems using limited vocabulary. The performance of these systems breaks down for a large vocabulary because of the confusability in the gross characteristics of many words. The confusability is increased because the fixed frame size parametrization fails to capture short duration changes. As an example, if centisecond processing is taken as standard, the parameters are obtained at ten millisecond intervals. But speech exhibits significant changes in as short a duration as one millisecond or lesser. Spectrogram of a segment of speech is shown in Fig. 2.2. The scale marks in the time axis are at ten millisecond intervals. Assuming scale marks at the center point of parameter sampling, one can see that such representations define the features at a gross level. Block data processing tends to smear the variations and gives an average behavior. Particularly, at transition regions, significant changes take place within ten milliseconds and these are not captured by such representations. On the other hand if we were to choose an arbitrarily fine resolution, then there will be too many details which increases variability in matching. Moreover number of data points become excessive which will increase the computation time. In the context

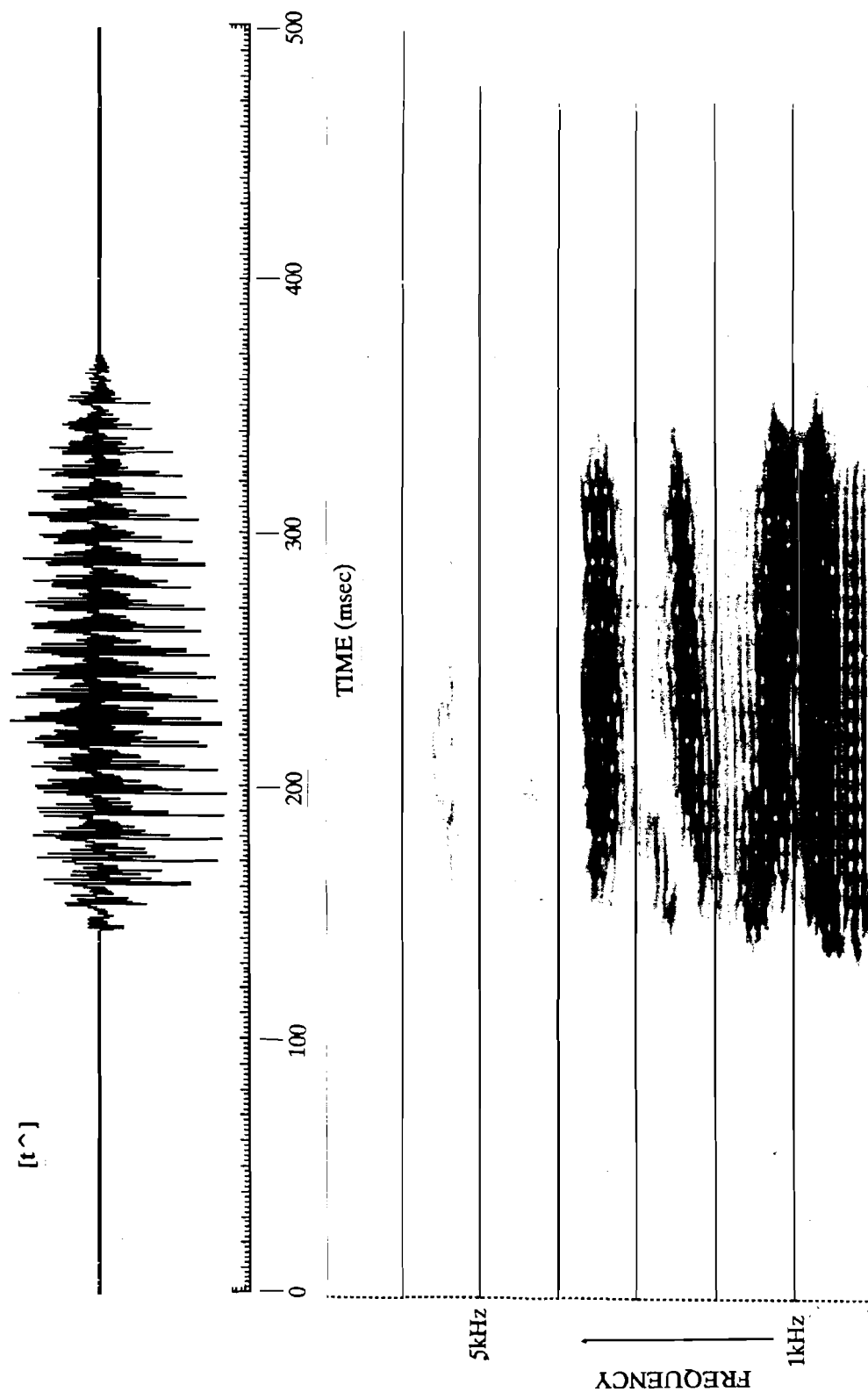


Fig. 2.2 Speech signal and spectrogram for a segment of utterance [ka]

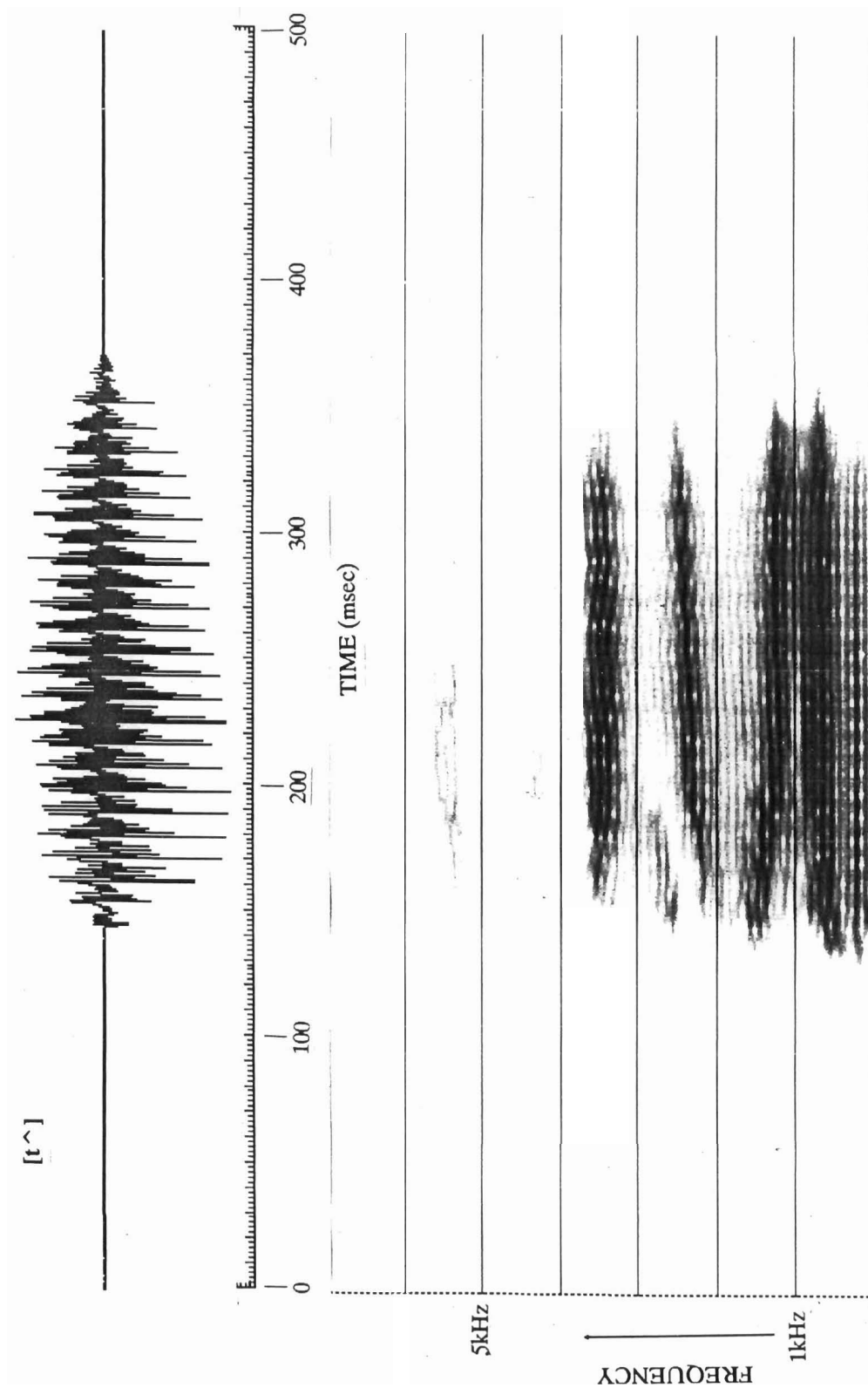


Fig. 2.2 : Speech signal and spectrogram for a segment of utterance [ka]



of speech excessive data poses a problem. Small variations in the parameter values add up in the difference score and the scores tend to be undesirably large even among close patterns. Another problem is that steady regions are unnecessarily processed with high temporal resolution. What is needed is a judicious choice of resolution dependent on the nature of the signal – regions exhibiting rapid variations in features processed with high resolution and steady state regions using low resolution. Some studies have used such varying resolutions by the use of trace segmentation [Lienard 1984] to improve the recognition performance.

The performance of any isolated word recognition scheme based on the parameter matching approach deteriorates as the size of the vocabulary increases. There are two aspects to this deterioration. One is in terms of the time taken for matching which increases almost linearly with the increase in the vocabulary size and the other is the increase in confusability due to many words having similar features. The different approaches to large vocabulary recognition are :

- a) a syntactic approach – where, at any time there exists only a small active vocabulary, and
- b) use of some form of classification to group similar sounding words and then use fine matching within the group.

The first technique uses context (language's syntactic context) to dictate what the active vocabulary is. The second approach is a two stage approach based on the similarity among the words in subsets of the vocabulary.

Use of fixed frame size for obtaining the features and the use of simple parameters as the features results in a transformation of the input signal into parameter sequences. Further processing is usually carried out on these

parameter sequences. This technique reduces the computation and interpretation complexities to manageable proportions but may lead to the loss of crucial information contained in the speech signal which cannot be derived by any amount of processing carried out on the parameter sequences.

In the context of a confusable vocabulary, it is this information which plays a key role in deciding the improvement in performance of a speech recognition system. This does not imply that one should derive all possible features before doing away with the signal. What is required is a judicious choice of parameters or features which is dictated by the context. The context defines the processing that is to be carried out. This suggests the use of knowledge based processing of speech signal to realize good performance in a recognition system.

## **2.2 KNOWLEDGE-BASED SIGNAL PROCESSING FOR SPEECH RECOGNITION**

The improvement in performance for the alpha-digit set by the use of signal dependent approach in the choice of parameters with little or no modification to DTW algorithm is reported in [Yegnanarayana 1984]. In the Indian language context the set of Hindi stop consonants is a highly confusable vocabulary and [Raman 1985] reports the improvement in performance by the use of signal dependent approach to dictate the choice of parameters, time resolution and the DTW algorithm. This work reports the results of recognition for individual groups of these consonants. The work also reports the use of language dependent constraints to dictate the strategy used for recognition. The work points out that for a confusable vocabulary the choice of processing strategy is crucial to obtain good recognition performance. The work reports that the phonetic nature of the language can be exploited by considering groups of the consonant set based on acoustic phonetic features.

The present work is aimed at addressing issues for achieving good recognition performance for utterances of the Hindi alphabet set. The vocabulary that is being considered here is a large one due to various combinations of consonants and vowels that are possible. The vocabulary is a highly confusable one too. The large and confusable vocabulary implies that standard approaches used in **IWSR** do not work for this vocabulary. Hence there is a need for a different approach to recognition. Search for a different approach to recognition in the context of a confusable vocabulary leads one to consider the nature of the vocabulary and the characteristics of the speech production mechanism in more detail.

Speech signal conveys a lot of information. It consists of sequence of regions with distinct features. Auditory perception of these leads to recognition of speech. In automatic speech recognition, we require a machine to do something similar or arrive close to it. Some of the features are clearly and conspicuously visible even in the waveform plot. Spectrogram, which is a pictorial representation of the spectral characteristics in speech as a function of time, clearly illustrates many of the features. Again, as in auditory perception, it is the human faculty of visual perception which makes it possible to discern the features in the spectrogram. Spectrogram (sometimes called visible speech) is used by phoneticians and linguists as the mainstay tool in the analysis of speech of a language. (The visible form makes it possible to carry out an objective analysis and arrive at quantitative measurement of feature characteristics.) Parametric representation of the speech signal by standard signal processing methods do not capture all the features that are present in the signal.

Speech signal is generated by a physical production mechanism. Physical constraints in the mechanism restrict the type of signals generated. Language imposes further restrictions. Only certain types or categories of sound are used. Speech sounds are therefore a subset of sounds producible by the speech

production mechanism. Signal processing can take advantage of this knowledge to make it more effective. In other words, knowledge driven signal processing can perform much better than conventional signal processing.

The required knowledge is obtainable from acoustic-phonetics which is the study of acoustic behavior of phonetic features in the speech of a language. By definition, phonetic features imply those speech features that have relevance from the point of view of the language and the conveyed message. Descriptions, features and characteristics of different speech sounds obtained from acoustic-phonetics are inherently speaker independent. Hence, with this approach, variability due to speaker characteristics can be handled in a systematic manner. This approach permits classification of speech sounds into groups and subgroups [Yegnanarayana 1991]. Such a classification reduces the problem space and makes it possible to carry out recognition in stages. In the context of isolated utterances of characters of Hindi, such an approach becomes necessary as it is essential to detect the occurrence of relevant and distinguishing features to obtain good recognition performance. In the next chapter, we discuss the use of acoustic-phonetic knowledge in arriving at the methodology for recognition of CV utterances.

## **ACOUSTIC-PHONETIC KNOWLEDGE OF CV UTTERANCES IN HINDI**

In this chapter we discuss the acoustic-phonetic knowledge for isolated utterances of characters of Hindi which embodies the relationship between acoustical properties of speech and the physiological mechanism of speech production in the utterances of characters. We also discuss how this acoustic-phonetic knowledge in the form of 'acoustic events is manifested as a set of parameters and features in the signal. At the end of this chapter, we discuss how this knowledge is used in arriving at the issues for study to realize recognition of utterances of isolated characters of Hindi.

### **3.1 INTRODUCTION TO ACOUSTIC-PHONETICS**

Acoustic-phonetic knowledge may be defined as 'that obtained from the study of acoustic behavior of phonetic features in the speech of a language. As stated in [Broad 1975] the fundamental concepts in acoustic-phonetics are :

1. Interpretation of the continuously varying acoustic parameters as a sequence of discrete units that can be related to linguistic structures,
2. The characterization of the relation of phonetic equivalence between speech segments.

The acoustic-phonetic descriptions use the articulatory state and their dynamics as a guide. The different acoustic sources in the speech production mechanism form an important basis for these descriptions as they dictate the different speech wave types. Visual features abstracted from the sound spectrogram are used as the elements for the description. Since speech has many landmarks, a major problem is one of deciding which landmark at which instant is the one to pay attention to. The choice of the visual features are the result of complex decisions on which features of the spectrogram are the most important.

With this background on what comprises acoustic-phonetics, let us now look at the Hindi alphabet which is the vocabulary of interest in our study.

### **3.2 THE HINDI ALPHABET**

Hindi language exhibits phonetic nature which is explicitly brought out in the organization of the alphabet (See Table-3.1). There are two broad categories in the alphabet – vowels and consonants. The vowels of Hindi include both diphthongal and non-diphthongal sounds. There is a phonological distinction between short and long forms of some of the vowel sounds which are represented by distinct symbols. Some long forms exhibit minor phonetic variations from their short forms.

Consonants in Hindi are subdivided into three groups – stop and nasal consonants, semivowels and fricatives. The structure within each group reflects different places of articulation and manners of production of consonants. Table-3.1 shows a rearranged version of the normal table which continues to retain the structure. The five columns in the stop and nasal group correspond to the five different places of articulation – velar, palatal, retroflex, dentalalveolar and bilabial. The rowwise arrangement reflects grouping based on the manner of production of the consonants. The four rows in stop consonants belong to the nonnasal

**Table-3.1: The Hindi alphabet arranged to reflect classification based on acoustic-phonetic features**

**VOWELS AND DIPHTHONGS**

Tongue hump position	Back	Back	Central/ Back	Front	Front	Central to Front	Back
Tongue hump height	Mid	Low	Low	Mid	High	Low to High	Low to Mid
Lips rounding	Yes	Yes	No	No	No	No	No to Yes
Length of vowel							
Short	[U]	—	[^]	—	[I]	—	—
Long	[u]	[o]	[A]	[e]	[i]	[aj]	[aw]
	VOWELS					DIPHTHONGS	

**SEMIVOWEL AND NASAL CONSONANTS**

Coupling of nasal tract by lowering of velum	Coupled	Decoupled
Sub group	Nasal	Semivowels

**NASALS**

Velar	Palatal	Retroflex	Dentalveolar	Bilabial
[g~]	[G]	[N]	[n]	[m]

**SEMIVOWELS**

PLACE OF ARTICULATORY RELEASE			
Palatal	Alveolar trill	Alveolar lateral	Labiodental
[j]	[r]	[l]	[v]

**FRICATIVE CONSONANTS**

PLACE OF ARTICULATORY RELEASE			
Glottal	Palatal	Retroflex	Alveolar
[h]	[s']	[S]	[ʃ]

**STOP CONSONANTS**

MANNER OF ARTICULATION		PLACE OF ARTICULATORY RELEASE				
		Velar	Palatal	Retroflex	Dentalveolar	Bilabial
UNVOICED	Unaspirated	[k]	[tʃ]	[ɖ]	[ʈ]	[p]
	Aspirated	[kʰ]	[tʃʰ]	[ɖʰ]	[ʈʰ]	[pʰ]
VOICED	Unaspirated	[g]	[dʒ]	[ɖ]	[ɖ]	[b]
	Aspirated	[gʰ]	[dʒʰ]	[ɖʰ]	[ɖʰ]	[bʰ]

Note : Phonetic symbols used are as in Computer Phonetic Alphabet [Lennig 84] with modification for aspiration as suggested in [Ganesen 75]

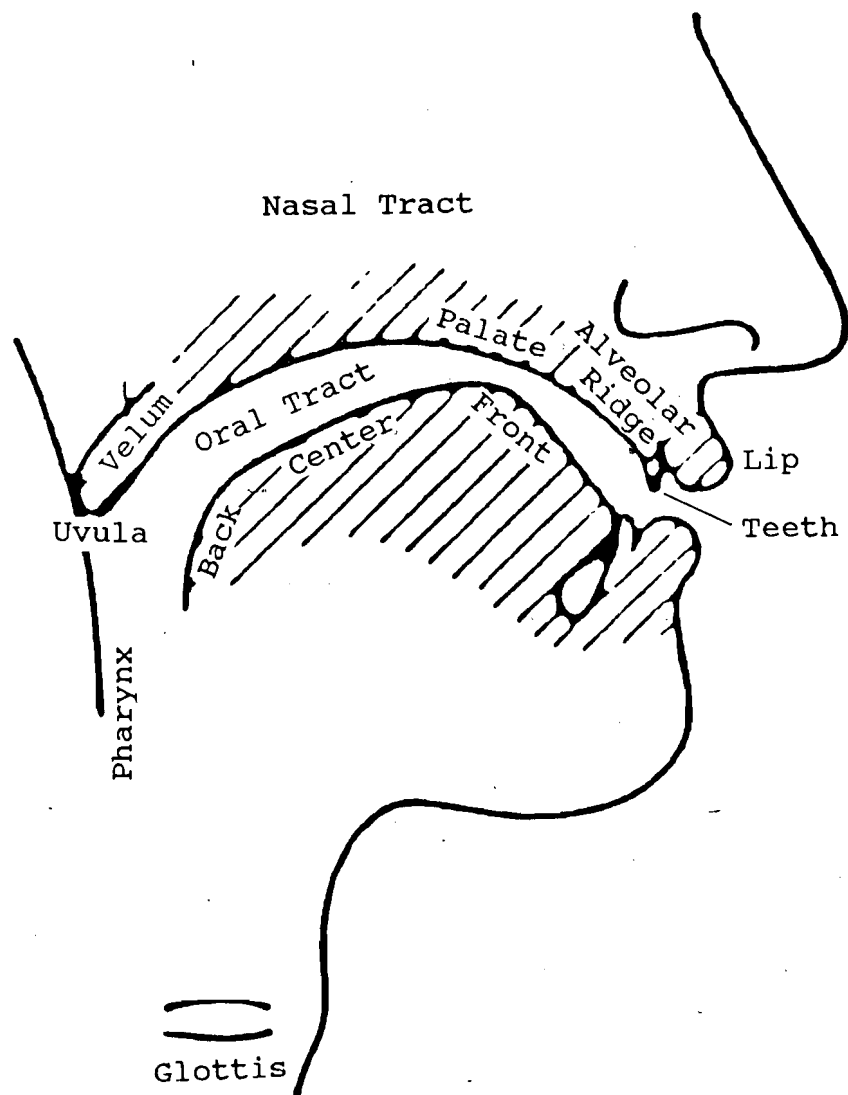
category. The nasal consonants are shown in a separate row. Stop consonants in the first and second rows belong to the unvoiced category. Glottal excitation is absent during the production of these consonants. The third and fourth rows represent the voiced consonants. In these cases glottal excitation starts prior to the consonant release. Consonants are also categorized on the basis of presence or absence of aspiration. Consonants in the first and third rows are unaspirated, whereas the consonants in the second and fourth rows are aspirated. Nasal consonants are by nature voiced and unaspirated. Semivowels too are by nature voiced and unaspirated, and different semivowels are differentiated based only on their place of articulation. The order of arrangement for semivowels is from the inner to the outer point of articulatory stricture in the mouth – palatal, alveolar (trill), alveolar (lateral) and labiodental. Fricatives in Hindi are always unvoiced. The order of arrangement for fricatives in the table is based on the point of stricture in the mouth – glottal, palatal, retroflex and alveolar.

Characters in Hindi are combinations of consonants (C) and vowels (V). They occur in the forms of C, V, CV, CCV and CCCV, where C is any consonant from the consonant group and V is any vowel from the vowel group. In this study we consider characters of the type V and CV only.

### **3.3 SPEECH PRODUCTION MECHANISM**

Phonetics deals with systematic study of human speech sounds in order to describe and classify them. Phonetic features are described usually in terms of the articulatory process (see Fig. 3.1). As speech sounds are characterized by both dynamic and static behavior of the articulators, the description of speech sounds is based on the articulatory positions, articulatory dynamics, resultant acoustic sources and the sequence of events.





**Fig. 3.1 : Sketch of the parts of human speech production mechanism**

Among the two broad classes of speech sounds (vowels and consonants), vowels are produced with a relatively free passage of air stream in the vocal tract with the vocal cords set into strong vibrations. The vibrations are setup by the air pressure difference across the vocal cords and this results in quasi-periodic impulses of air being injected into the vocal tract. This functions as the acoustic signal source with periodic nature and wide-band characteristics. The vocal tract, an acoustic tube, imposes its transmission characteristics on the acoustic signal emanating from the source and the resultant signal is radiated from the mouth. The different vowels are generated by altering the shape of the vocal tract. The quality of a vowel is mostly dictated by the positions of the tongue and lips as these introduce maximal and conspicuous change in the shape of the vocal tract. The velum too plays a role in that it determines whether or not the nasal tract is coupled to the oral tract.

Consonants are produced when the air stream passing through the vocal tract is obstructed in some way by the articulators. The closure position is followed by a release towards an open tract configuration of the following vowel. The consonants differ depending on the place where the obstruction takes place as well as the manner of articulation in the production of the consonant. The obstruction can occur in many ways. The acoustic signal source (excitation) during consonant production can also occur in different ways – the release of articulator can result in impulse excitation at the point of constriction; random noise like excitation can occur when articulators are brought close together leading to turbulent air flow; and vocal cord vibrations can occur along with articulatory release leading to periodic impulse excitation. The characteristics of the speech production mechanism and consequent acoustic features during consonant production in CV utterances of Hindi are discussed in a later section in this chapter.

### 3.4 NATURE OF SPEECH SIGNAL IN CV UTTERANCES

Typical speech waveform and its characteristics are illustrated in Fig. 3.2 taking the example of two utterances [k<sup>h</sup>ʌ] and [tʌl]. The figure highlights the details in the consonantal parts of the utterance. The initial burst, frication and aspiration characteristics of the consonant, and the voicing characteristics in the initial part of the vowel are shown. The complex nature of the speech signal and its time varying characteristics are clearly seen from the figure.

The burst is a very short duration signal with an amplitude significantly large relative to the neighboring regions of the signal. In the illustrated example, there are several cycles of the signal within this short duration which indicates high frequency concentration. The frication region which is observed in the initial part of utterance [tʌl] also exhibits rapid oscillations in the signal, but for this case, the signal exhibits randomness and is present for a significantly longer duration. The amplitude characteristics in this region are also distinctly different from that of the burst. The aspiration region in the signal shows mixed nature. There is random large amplitude variation with random rapid variations superimposed in it. The vowel region in both the utterances clearly show periodicity in the signal. This can be seen from the repetitive nature of the waveform. The time period of this repetition is called the pitch period. The nature of the waveform within a pitch period exhibits damped sinusoidal behaviour. This reflects the resonance behavior of the speech production mechanism. The spectral characteristic of this resonance behavior is the formant structure. The resonance behavior shows distinctly different characteristics for the two examples shown. The illustration shows these features clearly, but the features exhibit large variability dependent on the context of the consonant and vowel.

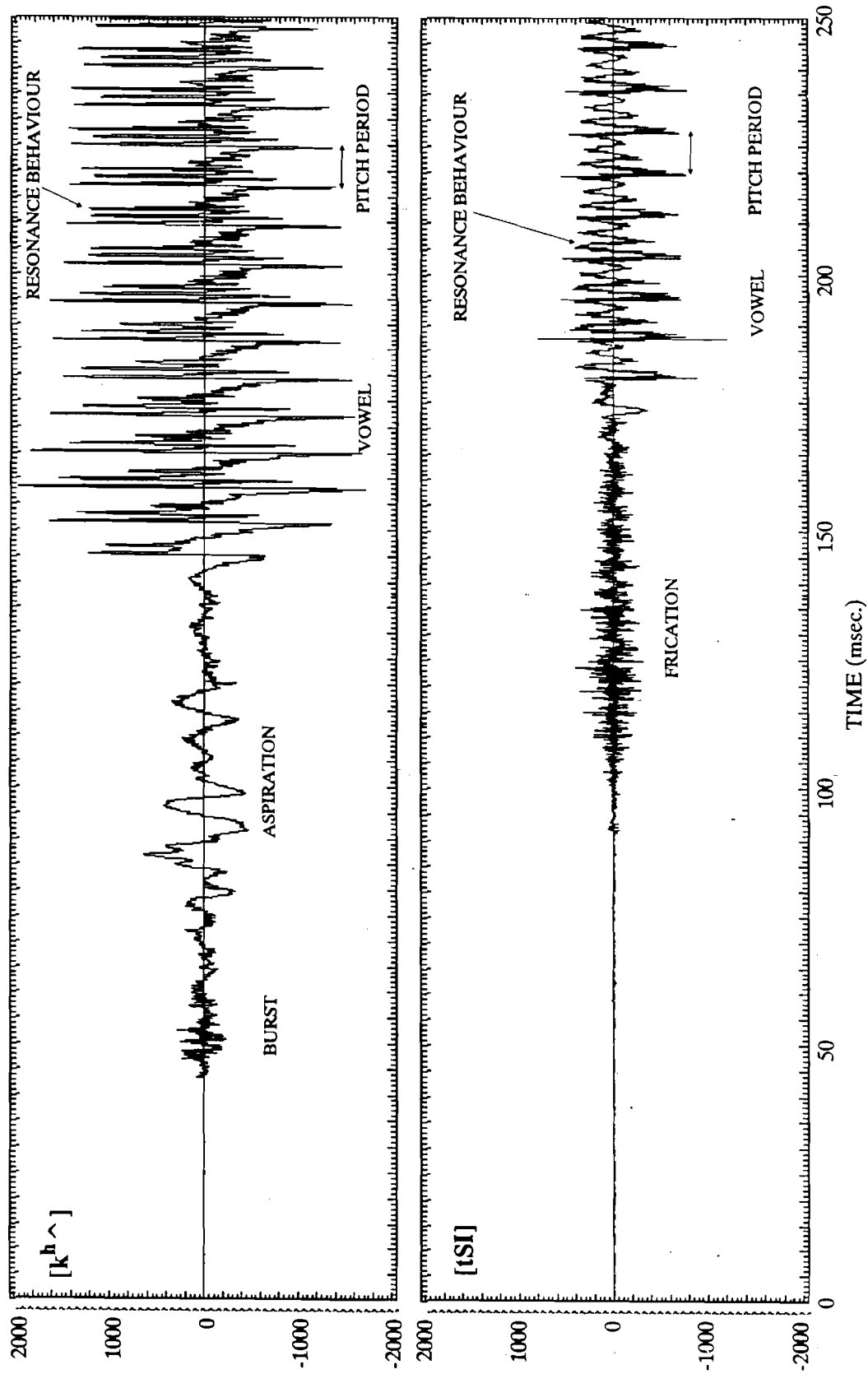


Fig. 3.2 : Speech waveform for the utterance  $[k^h^~]$  and  $[tSI]$

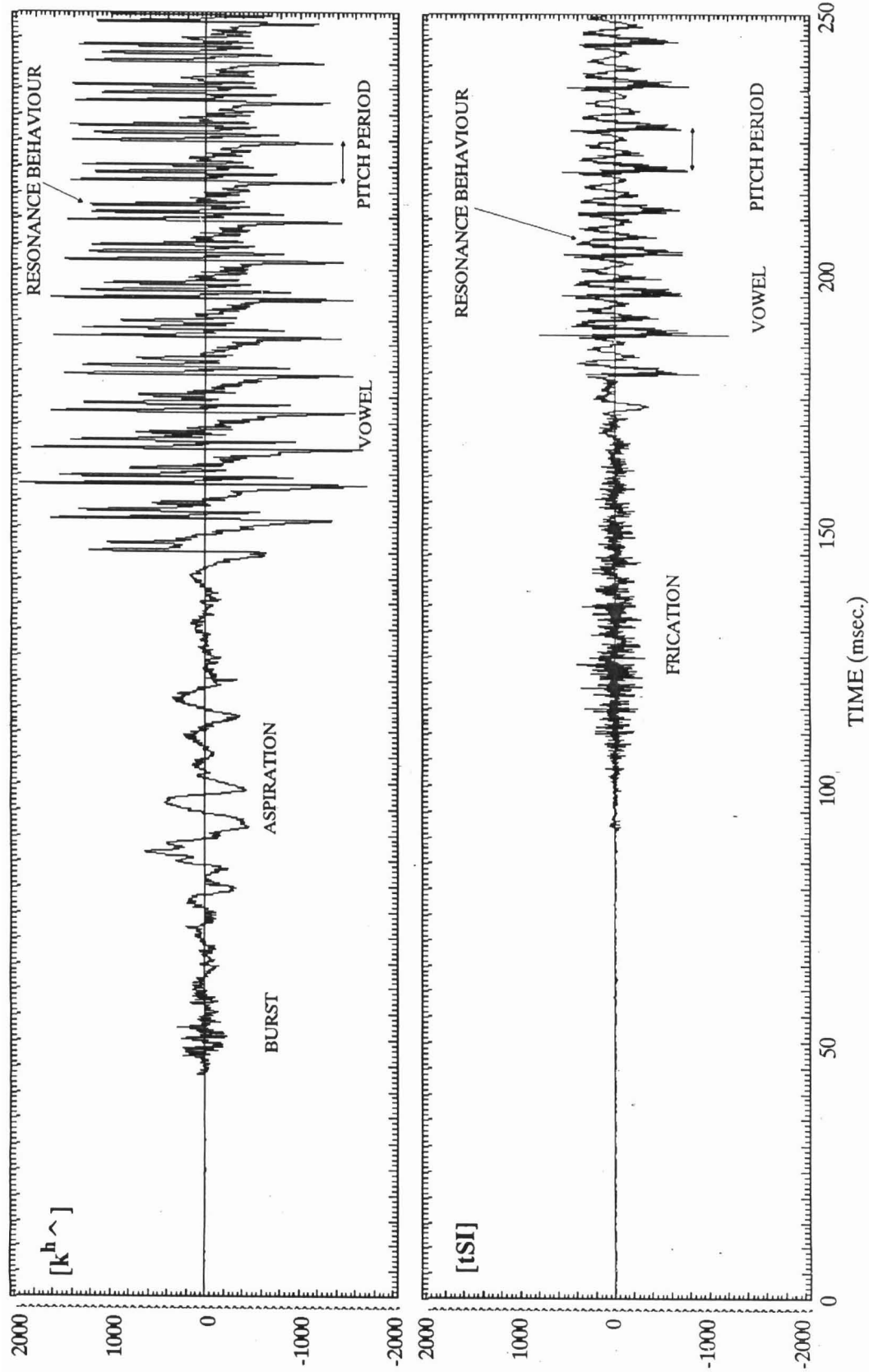


Fig. 3.2 : Speech waveform for the utterance  $[k^h ^]$  and  $[t^h SI]$

### **3.5 ACOUSTIC FEATURES OF CV UTTERANCES**

The Hindi CV utterances have an independence in the description of the consonantal part and the vowel part which makes it possible to combine any consonant with any vowel to form an utterance. The consonant part of an utterance is one of thirty three possible consonants of Hindi (thirty four if we consider the case of an utterance without any consonant), while, the vowel part is one of ten different vowels/diphthongs. The acoustic characteristics of the CV utterance set show distinct commonality in the features among groups of the CV utterance set. The common features are due to the similarity in the way the consonant is produced. This is usually termed the manner of articulation of the consonant. Individual consonants within a group are distinguished by the specific place of articulatory release (or place of articulation).

#### **3.5.1 Features of broad classes and regions in CV utterances**

The CV utterances can be grouped into eight broad classes based on the manner of production of the consonants (or rather the consonant category). This grouping follows directly from the phonetic ordering in the Hindi alphabet. The eight different categories (including the case of vowel only) are indicated in Table-3.2.

Utterances in each of these categories exhibit different regions with distinct characteristics. These regions show variations for the different cases of utterance mostly due to the context of the vowel region in the utterance and the place of consonantal stricture in the vocal tract. Table-3.3 identifies the different regions that occur in the various categories of utterances. The asterisks indicate the presence of a region type in a category and the order (from left to right) of the regions indicated as being present is the sequence of their occurrence in the utterance.

**Table-3.2 :** Broad classes of Hindi CV utterances based on the manner of production of consonant

1. Unvoiced unaspirated stops
2. Unvoiced aspirated stops
3. Voiced unaspirated stops
4. Unvoiced aspirated stops
5. Nasals
6. Semivowels
7. Fricatives
8. No consonant

**Table-3.3:** Significant regions in broad classes of utterances

BROAD CLASS	REGION							
	VNG	RLT	ASP	VAS	NSL	SMV	FRC	VWL
UV-UA-SC		*						*
UA-A-SC		*	*					*
V-UA-SC	•	•						*
V-A-SC	*	*		*				*
NSLC					*			*
SMVC						*		*
FRCC							*	*
VWLO								*

VNG : Voicing

RLT : Stop release transient

ASP : Aspiration

VAS : Voiced-aspiration

NSL : Nasal

SMV : Semivowel

FRC : Fricative

VWL : Vowel

All regions, other than the stop release transient, occur for a significant duration in the utterance. The region to the left of the vowel-like region in the utterance can be considered as the consonantal part. Thus the different region types that can occur in a CV utterance of Hindi are:

- a) Voicing (VNG) – that preceding the stop release in voiced stops
- b) Stop release transient (RLT) – burst
- c) Aspiration (ASP)
- d) Voiced–aspiration (VAS)
- e) Nasal (NSL)
- f) Semivowel (SMV)
- g) Frication (FRC)
- h) Vowel-like (VWL)
- i) Silence/background noise (SIL)

Region type (SIL) is included for the sake of completeness and represents region prior to utterance start as well as region after the utterance end. The regions are demarcated by specific events which trigger the change. The significant instants of change in an utterance are:

- a) Start of utterance (ST)
- b) Articulatory release instant (AR)**
- c) Initiation of glottal vibrations (GI)
- d) Vowel onset instant (VO)
- e) Termination of glottal vibrations (GT)
- f) End of the utterance (ND)

These instants in relationship to the occurrence of different regions in an utterance are indicated in Table-3.4. It can be observed that there is co-occurrence of some of these instants in different utterance groups.



**Table 3.4 :** Time sequence of instants and regions in broad classes of utterances

<b>UV-UA-SC</b>	
<b>UV-A-SC</b>	
<b>V-UA-SC</b>	
<b>V-A-SC</b>	
<b>NSLC</b>	
<b>SMVC</b>	
<b>FRCC</b>	
<b>VWLO</b>	

The basis for the formation of groups, region types and instants is from acoustic-phonetics and is similar to segment type features discussed by [Fant 1973]. This form of grouping of utterances and tabulation of different regions in an utterance suggest that the first level discrimination in terms of the group to which an unknown utterance belongs is possible only if we can detect the regions that occur. We shall now look at the typical acoustic features and the parametric behavior of the different region types.

### **3.5.2 Acoustic features of regions**

The following subsections discuss the production mechanism characteristics and the acoustic features in the different regions of CV utterances.

#### **3.5.2.1 Voicing (VNG)**

The voicing signal is produced when glottal vibrations are present with a complete closure of the vocal tract. Though glottal vibrations form wide band excitation, the absence of a radiation point causes the glottal vibrations to be heavily filtered and attenuated. The signal appears to have only the pitch frequency component. The signal is periodic and nearly sinusoidal. The signal energy is low but remains steady with very little fluctuations. There is no spectral structure due to the absence of higher frequency components. The absence of transmission through vocal tract also implies an absence of formant structure in the signal. Consequently, neither the shape of the vocal tract nor the place of articulation has any significant influence on the signal characteristics in this region. Though sustainable for a significant duration, the voicing signal usually exhibits a decrease in the amplitude near the end of the region. This is attributable to the reduction in glottal vibration strength as a result of decrease in the air flow across the glottis due to complete closure.

### 3.5.2.2 Stop Release Transient (RLT)

This region is characterized by signal of very short duration. The significant excitation is that due to the abrupt pressure release causing an impulse excitation at the place corresponding to the point of closure in the vocal tract. The nature of the excitation signal is dependent on the release dynamics and is a function of the place at which the articulatory release occurs. It is also a function of the nature of the preceding and following regions. Since the vocal tract imposes its transmission properties as well, the stop release signal exhibits differences in characteristics due to the shape of the vocal tract determined by the following vowel context. Thus there are many aspects to the differences in characteristics of the stop release transient. We shall now consider each of these aspects in more detail.

In voiced stops, since glottal vibrations persist right through the release, the release transient signal is embedded in voicing signal. The presence of glottal vibrations during closure, leads to lower pressure build-up when compared to unvoiced stops. As a consequence, the release transient in voiced stops has a lower amplitude.

In aspirated stops, the pressure build-up during closure is larger than in unaspirated stops. The production of aspiration calls for significantly larger air flow and the larger pressure build-up occurs in anticipation of this. This leads to a very large amplitude signal at release.

The stop release as a function of 'place' shows distinct differences for the different cases. The differences arise mostly due to the articulatory release dynamics. The velar stop release has one or more epochs (impulse / shock excitation) and is conditioned by almost all of the vocal tract as it is at an inner point. Since the position is close to the glottis, it influences the start of glottal

vibrations after the release by delaying it. The palatal position has a large area of contact. The release occurs with a short duration of narrow opening. This causes turbulence leading to fricative like signal with decay in amplitude. The retroflex, denti-alveolar and bilabial stops have the closure near the mouth. Consequently, the release transient is not affected significantly by the vocal tract. The transient occurs for a very short duration.

### **3.5.2.3 Aspiration (ASP)**

The aspiration signal is produced with narrow opening in the glottis. The narrowness of the glottal opening causes an absence of glottal vibrations, but, is narrow enough to cause the air release at high velocity. The acoustic signal source is the turbulence arising out of the high velocity air flow at the constriction as well as that due to any obstruction in the direct path of the high velocity air flow in the supraglottal region. The turbulence forms the wideband acoustic excitation of the vocal tract. It is characterized by randomness in both time and amplitude of the epochs (impulse excitation). Since the excitation occurs at an inner point in the vocal tract, it is subjected to the transmission (filtering) properties of all of the tract. Since the aspiration is always preceded by stop release in Hindi CV utterances, the air flow is initially very large. Consequently the excitations exhibit large energy at the beginning of the region which decays with time. The resultant signal exhibits large amplitude with random fluctuations and decay towards the end of the region. The turbulence in the initial part of aspiration occurs near the place of stop stricture rather than at glottis. In the production of certain vowels, the vocal tract exhibits constriction at some position which is narrower than the glottal constriction. In these cases, the acoustic signal source is at a posterior point in the vocal tract and not at glottis. In either case, the resultant signal exhibits spectral structure which is a function of the shape formed by the vocal tract in anticipation of the following vowel. As a consequence of the preceding stop release, vocal tract shape change occurs in the initial part of aspiration. This results in changes in the spectral structure in the initial part.

#### **3.5.2.4 Voiced–Aspiration (VAS)**

Voiced–aspiration is also produced with narrow opening of glottis, but, along with the high velocity air flow, the glottis is set into vibrations. Thus the acoustic signal source consists of two components, one is the random excitation due to turbulence and the other is the periodic excitation due to glottal vibrations. The glottal vibrations are relatively weak because of incomplete closure. The voiced–aspiration is also preceded by the stop release in Hindi CV utterances and this causes the air flow to be initially very large. A spectral structure is imposed on the mixed excitation in transmission through the vocal tract. The initial part exhibits change in spectral structure because of the preceding stop release. The resultant speech signal exhibits periodic waveform mixed with random fluctuations.

#### **3.5.2.5 Nasal (NSL)**

Nasal consonants are produced with a complete closure at some position in the vocal tract but with the velum lowered. This results in the nasal tract being coupled to the vocal tract system. Since the velum is inwards from the position of closure, the nasal tract provides a free path for the air flow. Thus there is no pressure build–up at the closure. The closure is subsequently released to realize an open tract configuration of the following vowel in CV utterances. The acoustic signal source is the glottal vibrations which is initiated prior to the release of the closure in the oral tract. Prior to the articulatory release, the acoustic signal passes through the nasal tract and is radiated from the nostrils. Thus the radiated signal is characterized by the nasal tract transmission properties. The anterior part of the oral tract up to the closure becomes an acoustic chamber coupled to the acoustic tube. The coupled acoustic chamber functions as a frequency selective energy absorber. Hence, the spectral structure of the radiated signal exhibits nulls (significant reduction in energy) at the absorption frequencies. The signal energy is distinctly lower than that after release though the relative strength of glottal vibrations remains nearly at the same level. Since there is an absence of pressure build–up at the closure, the articulatory release does not induce a new

acoustic signal source. The effect of articulatory release is merely to alter the structure and shape of the acoustic tube and consequently its transmission properties.

#### **3.5.2.6 Semivowel (SMV)**

Semivowels are produced when the articulators are brought in contact at some place in the vocal tract but forming an incomplete closure. There is thus a clear path for unobstructed air flow in the vocal tract even during closure. Strong glottal vibrations form the acoustic signal source. The distinct feature is that arising out of the specific shape of the vocal tract dictated by the nature of closure. This is reflected in the spectral structure imposed on the periodic impulse acoustic signal generated from glottal vibrations. Here again, as in nasals, the articulatory release does not induce a new acoustic source. The vocal tract is opened out to the following vowel's configuration. The signal before release exhibits distinctly lower energy than that after release.

#### **3.5.2.7 Frication (FRC)**

Fricative consonants are produced when the articulators are brought close together creating a narrow constriction at some point in the vocal tract. The narrow constriction partially obstructs the air flow causing significant increase in its velocity. The high velocity air flow leaving the constriction results in turbulence when striking an articulatory obstruction in its direct path. The acoustic signal source is this turbulence which is random noise like with high frequency components as dominant. The random noise like excitation is subjected to the transmission characteristics of the vocal tract acoustic tube mostly by the part outwards from the point of constriction. But for the case of glottal frication, the constriction in the vocal tract for all other fricatives is close to the mouth (radiation point). Hence the resultant random noise like signal is subjected to minimal filtering and has mostly high frequency components. The glottal frication exhibits a filtered version of the random noise like signal because of transmission through the

complete length of vocal tract. The glottal frication is produced similar to aspiration. But, there is a major difference. Aspiration is always preceded by a stop release whereas, glottal frication starts from silence.

#### **3.5.2.8 Vowel-like (VWL)**

The vowel-like region is the region after vowel onset in CV utterances. The vocal tract is relatively open and hence causes free air flow. Strong glottal vibrations result in periodic wide band impulse excitation of the vocal tract and this forms the acoustic signal source. Spectral structure is imposed by the transmission characteristics of the relatively free vocal tract acoustic passage. The signal exhibits high energy level due to the open vocal tract configuration. The diphthongal vowels of Hindi are produced with a gradual change in the vocal tract shape over this region. Nasal utterances usually exhibit nasalization in the vowel region as well. The radiated signal in this case is that due to passage through nasal tract and the oral tract. For most cases of CV utterances except the aspirated and voiced-aspirated class, the initial part of the vowel-like region exhibits vocal tract shape change due to the preceding consonantal release. The initial shape is that due to the constriction at the place of articulation as well as the anticipatory positioning for the following vowel context. The shape of the vocal tract, after the articulatory movement due to release is complete, is that of the vowel context. Thus the beginning of this region exhibits changing or transitional characteristics.

### **3.6 METHODOLOGY FOR RECOGNITION BASED ON ACOUSTIC-PHONETICS**

From the discussion in the previous section, it is clear that the meaningful sounds that we produce are well defined. Sequences in which these sounds are produced are also unique. But the sounds do not conform to assumptions made in the usual source-system model of speech production for analysis purposes. Some of the assumptions in this model are :

- An all pole model for the system
- Short-time invariance of the model
- Simplified assumption about the source
- The source and the system features are independent

Some of the problems that are not handled well by this model are the following :

- Dynamic behavior of the source and system
- Interaction between the system and source
- Coupling between the vocal tract and nasal tract
- Excitation at different points in the vocal tract system
- Multiple excitation

In most cases the information is available in the signal and is visible (and perceivable) in the temporal and spectral domains as in the signal waveform and spectrogram, respectively. The problem is lack of suitable signal processing techniques to extract the relevant information.

As evinced from the discussions in the previous section the CV units are dynamic sounds. There are distinct regions in the utterance with characteristic features. This suggests that by considering different regions of the CV units independently, it is possible to carry out appropriate processing dictated by the nature of the region. The three important regions for CV utterances are the region before vowel onset instant, the transition region immediately following the articulatory release instant and the steady vowel region. It is necessary to identify the instants where these major acoustic events take place. Specifically, it is necessary to determine the start of utterance, the vowel onset instant and the articulatory release instant. These instants help to identify the consonantal region,

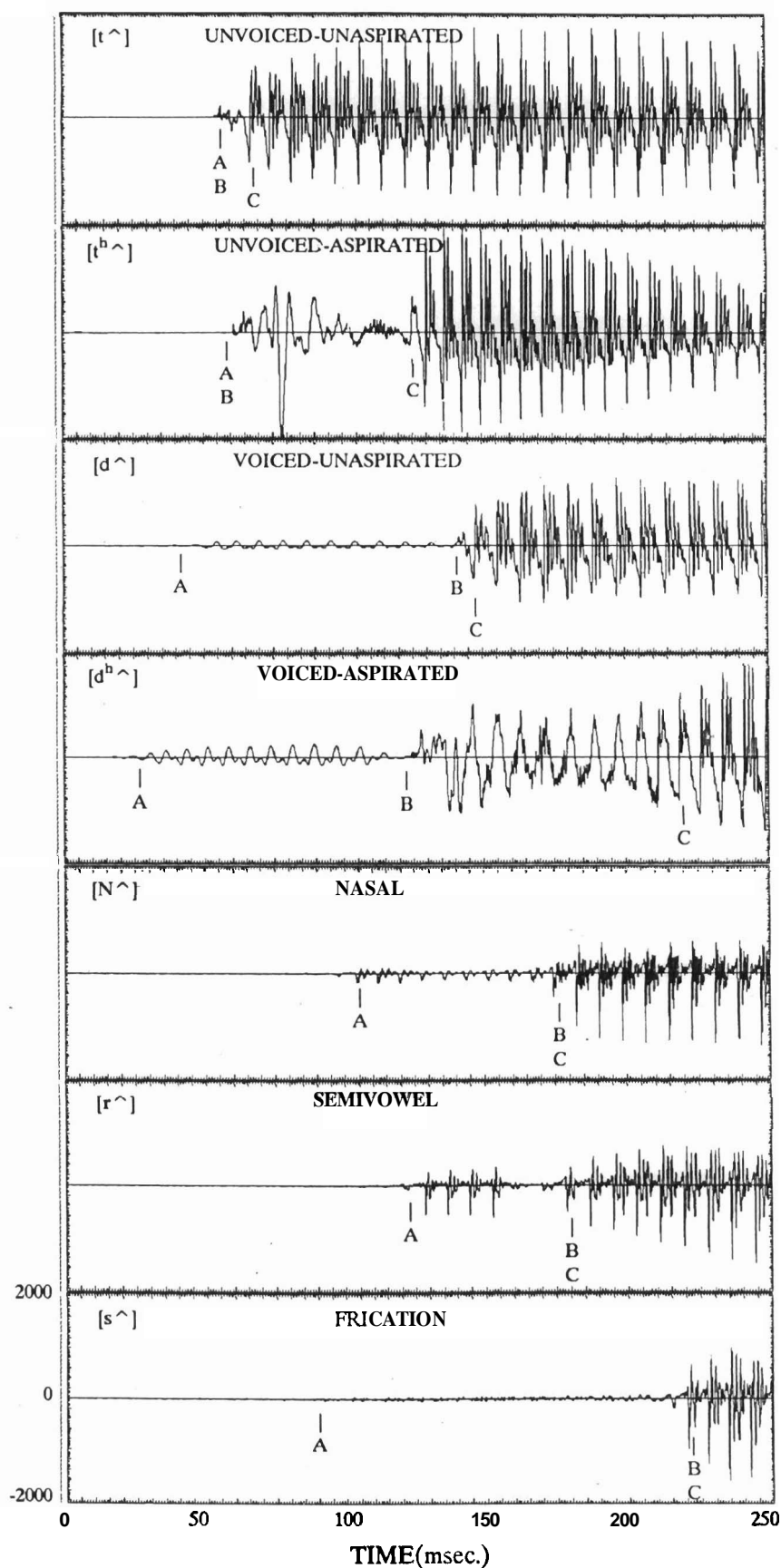


the transition region and the vowel region. These regions for the different manners of consonants are shown in Fig. 3.3. Detection of these instants and regions is the first issue that needs to be addressed.

The different consonantal classes have distinctly differing gross characteristics in the nature of the speech signal in the consonantal region, which is the region before vowel onset. Hence, if we confine attention to this region, it is possible to use relevant and appropriate parameters to discriminate amongst the different consonantal classes. The assumption is that the region for analysis is known apriori. The speech signal in this region should be processed to obtain features relevant for determination of the consonant class (manner), namely, unvoiced–unaspirated, voiced, aspirated, voiced–aspirated, nasal, semivowel and fricative.

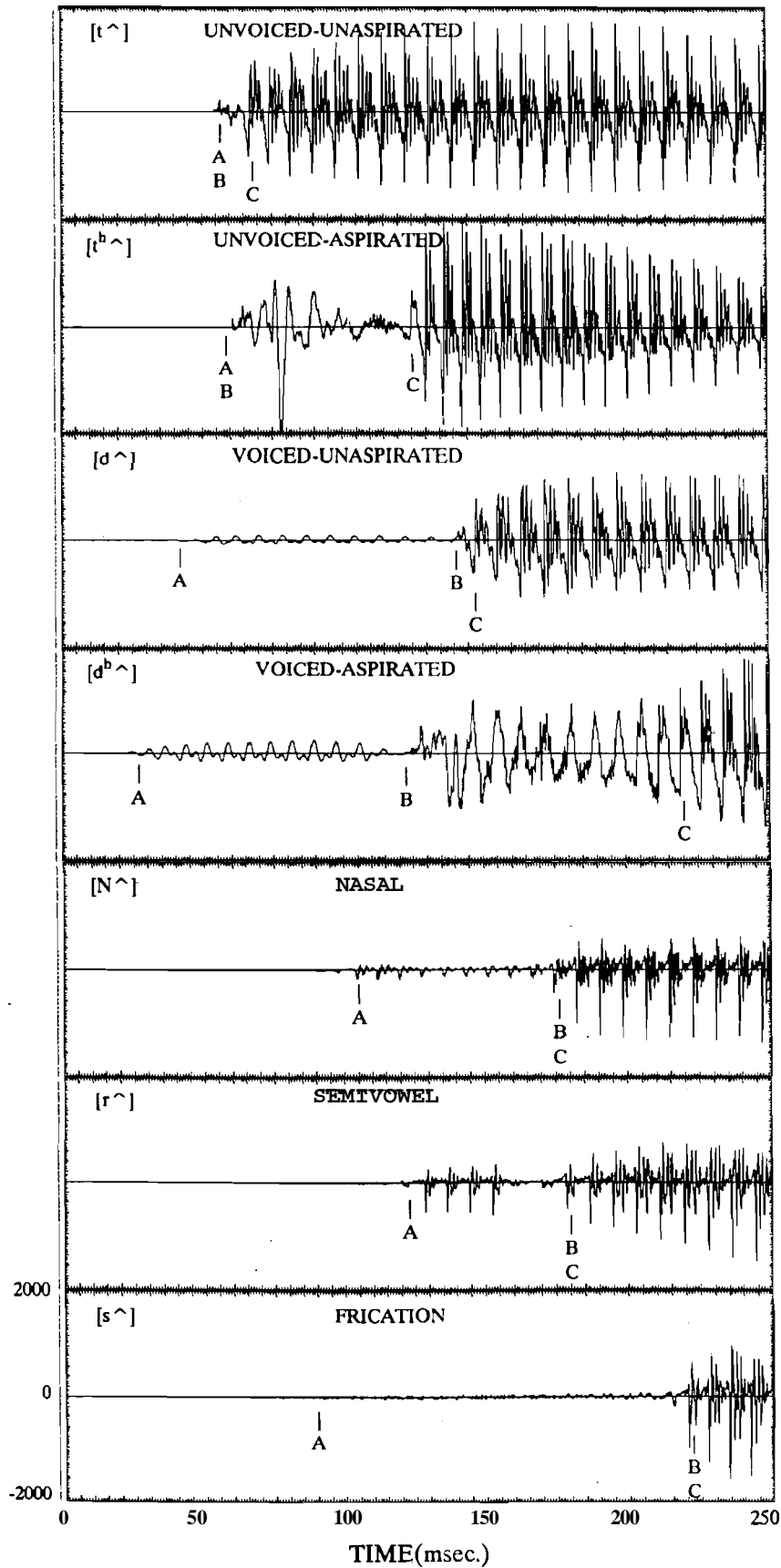
The vowel region needs to be processed to determine the vowel type of the **CV** utterance. For monophthongs it is enough to use the signal characteristics in the middle of the vowel region. For diphthongs, it is necessary to determine the nature of vocal tract shape change that occurs in the vowel region. The spectral features in the vowel region are dictated by the specific shape of the vocal tract besides its size, which is variable from speaker to speaker. But the vowel type is dictated more by the gross shape rather than the specific dimensions of the vocal tract. Therefore, for speaker independent recognition, the signal in this region is processed to extract the information of the vocal tract shape through analysis of the spectral features.

The cues for the place of articulation in **CV** utterances are in the signal prior to as well as during the articulatory release. For cases of consonantal regions with significant signal strength, the spectral characteristics of the signal in the consonantal region is determined by the nature of acoustic excitation and the extent of vocal tract that comes into play. As an example, let us consider the



**Fig. 3.3 : Significant instants and regions for the different manners of consonants**

A: Start of utterance; B: Articulatory release; C: Vowel onset instant



**Fig. 3.3 : Significant instants and regions for the different manners of consonants**

A: Start of utterance; B: Articulatory release; C: Vowel onset instant

iodic excitation due to glottal vibrations even during the consonant production. But the oral tract exhibits a complete closure and, the nasal tract provides the path for air flow. The glottal excitation is conditioned (filtered) both by the inner cavity formed due to the closure in the oral tract and the nasal passage. Thus the spectral characteristics of the resultant signal is conditioned by the place where the closure occurs in the oral tract (in other words, the place of articulation). In the case of stop consonants, the articulatory release results in a transient signal which is present for a very short duration. The nature of the transient signal is a function of the place of articulation, the manner of the consonant (consonant class) and the anticipatory vocal tract position due to the following vowel context. But the signal is generally weak prior to release and this causes difficulties in extracting the spectral characteristics reliably.

The transition region which is a short (30 msec.) duration immediately following the articulatory release holds the important cue to determine the place of articulation. This is because the dynamic characteristics of the vocal tract shape change is determined by the place of articulation and the following vowel context. For most of the consonant classes this occurs along with vowel onset. Since the vowel region exhibits relatively high energy signal, the signal is amenable for detailed spectral characterization. The exceptions are the aspiration and voiced-aspiration classes. For these cases, the articulatory release is immediately followed by aspiration or voiced aspiration rather than vowel onset. Therefore, the transition region occurs during the initial part of aspiration and voiced-aspiration respectively. Thus, characteristics of the transition region can be used reliably for determination of the place of articulation.

Having evolved methodology for carrying out different aspects of discrimination in different parts of the CV utterances, the overall recognition of the CV unit appears to be trivial, calling only for combining the results. But, this is true only if we are in a position to realize cent percent correct discrimination of

these different parts. Even if cent percent correct discrimination is not realizable, it is possible to select both best and second best choice in different parts and use the consistency among the choices to establish the identity of the overall CV utterance.

In summary, the recognition of the CV utterance can be looked upon as four separate problems.

1. Determination of significant instants namely, start of utterance, articulatory release instant and vowel onset instant. These three instants are needed to identify different parts for the individual analysis,
2. Determination of consonant manner by analysing the region before vowel onset instant,
3. Determination of vowel type of the utterance by processing the region in the utterance after vowel onset, and,
4. Determination of the place of articulation by analysing the region immediately following articulatory release instant to determine the nature of the transition.

Since the aim of the present study is to address the issues and arrive at an appropriate methodology for recognition, each of the above indicated aspects of discrimination is considered independently as a study. If information from one is needed in the other, that information is presumed to be known and the study is carried out. An important input needed in each of the studies is the instants separating the different regions for analysis. This is presumed available and is obtained through manual analysis of the speech data. Where appropriate and necessary, semi-automatic methods (specially processed parameters,) have also been used to identify these instants. Note that this detailed analysis is needed

only during training phase of a recognition system. Once trained, the classifier can be used to spot the required instants and regions automatically. The detection of significant instants is taken up as a separate study by itself. This forms the topic of the next chapter.

## **DETECTION OF SIGNIFICANT INSTANTS FOR IDENTIFICATION OF DIFFERENT REGIONS**

Speech signal for CV utterances exhibits regions with relatively steady characteristics as well as regions with rapidly varying characteristics. Important cues for the recognition of an utterance are present in both types of regions for different aspects of recognition. As an example, the quality of the vowel in a CV utterance is determined by the characteristics in the relatively steady region of the utterance; whereas for the consonantal part, the important cues for recognition lie in the region exhibiting transitional (rapidly varying) characteristics. Hence, for recognition of a CV utterance, there is a need to identify the regions with different characteristics in the utterance where detailed analysis needs to be carried out. The need is more so for the CV utterances of Hindi where, there exist a number of candidate utterances with distinguishing characteristics only in the consonantal part. As an example, consider the different characters of Hindi in a row of the stop consonants shown in Table-3.1 (Chapter 3). The first row identified as unvoiced-unaspirated stops, has the consonantal part of the character being different only in the place at which the articulatory release occurs. This is true for the other rows in the stop group, where the manner of production of the consonant is the same and only the place of stop release is different.

In this chapter, we shall first look at the instants separating the regions of interest when considering CV utterances. The characteristic features of the events at these instants are discussed, and a method to determine them is evolved. It is difficult to determine these instants from the signal waveform or from

the parameter contours derived from the signal (Fig. 4.1). Since these events for many cases of CV utterances correspond to the instants of significant excitation in the signal, special signal processing methods to determine the epochs have been used: Epochs determined using the above approach aid in the manual identification of the significant instants in CV utterances. These manually identified instants are used in different recognition studies. Although this method makes it possible to determine these instants with high positional accuracy, it is necessary to perform post processing to remove the unwanted epochs that are detected. Strictly speaking, the high positional resolution provided by the epoch detection method is not warranted. The computational burden is also high in this method. In the latter part of this Chapter, use of gross parameters along with neural network structures is studied for the detection of these significant instants.

#### **4.1 SIGNIFICANT INSTANTS IN CV UTTERANCES**

As Chapter 3, important events that demarcate different regions in any CV utterance are :

- a) start of the utterance,
- b) articulatory release instant, and
- c) vowel onset instant.

These instants occur in all CV utterances. In some utterances these instants co-occur. The start and end of an utterance identify the relevant span of an utterance for processing. The consonantal (C) part exhibits rapid movement of articulators after the articulatory release instant. Identification of the articulatory release instant is necessary for the determination of place of articulation. The vowel onset instant identifies the instant when strong vocal cord vibrations with relatively open tract sets in. Region prior to this instant may be considered as the



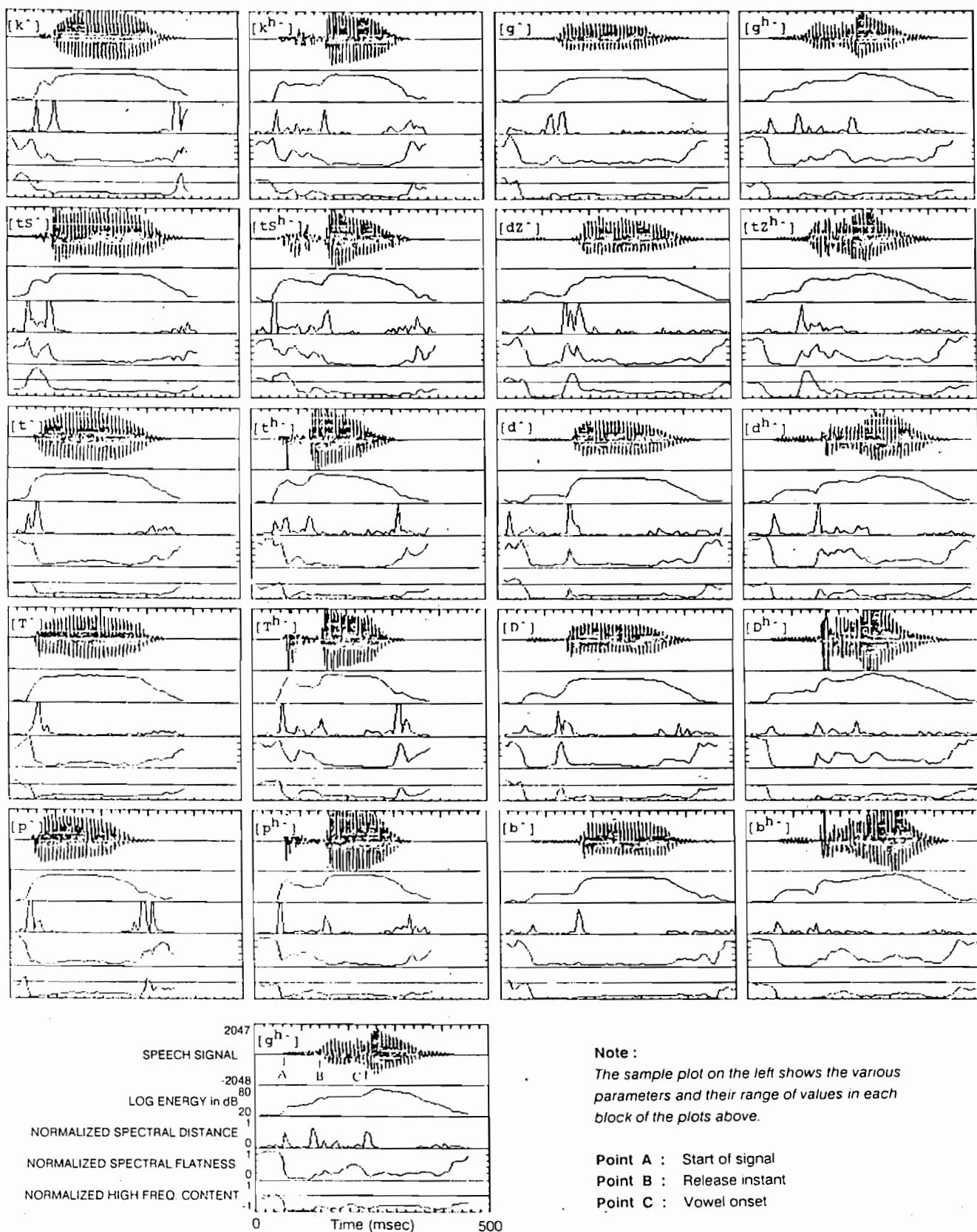
consonantal part of the CV utterance. The instants of initiation and termination of glottal vibrations play a role in describing the utterance type.

It is interesting to note that by focussing on these significant events, the emphasis on beginning and end of an utterance is not critical as in the case of IWSR. Using the three instants – start of the utterance, articulatory release instant and vowel onset instant – as reference, it is possible to determine the identity of a CV utterance without the need to carry out time normalization. Another benefit that is accrued is that the analysis methods can be tuned considering the region of interest and the nature of discrimination. The signal processing burden is also reduced since detailed analysis is carried out only in selected regions of the signal.

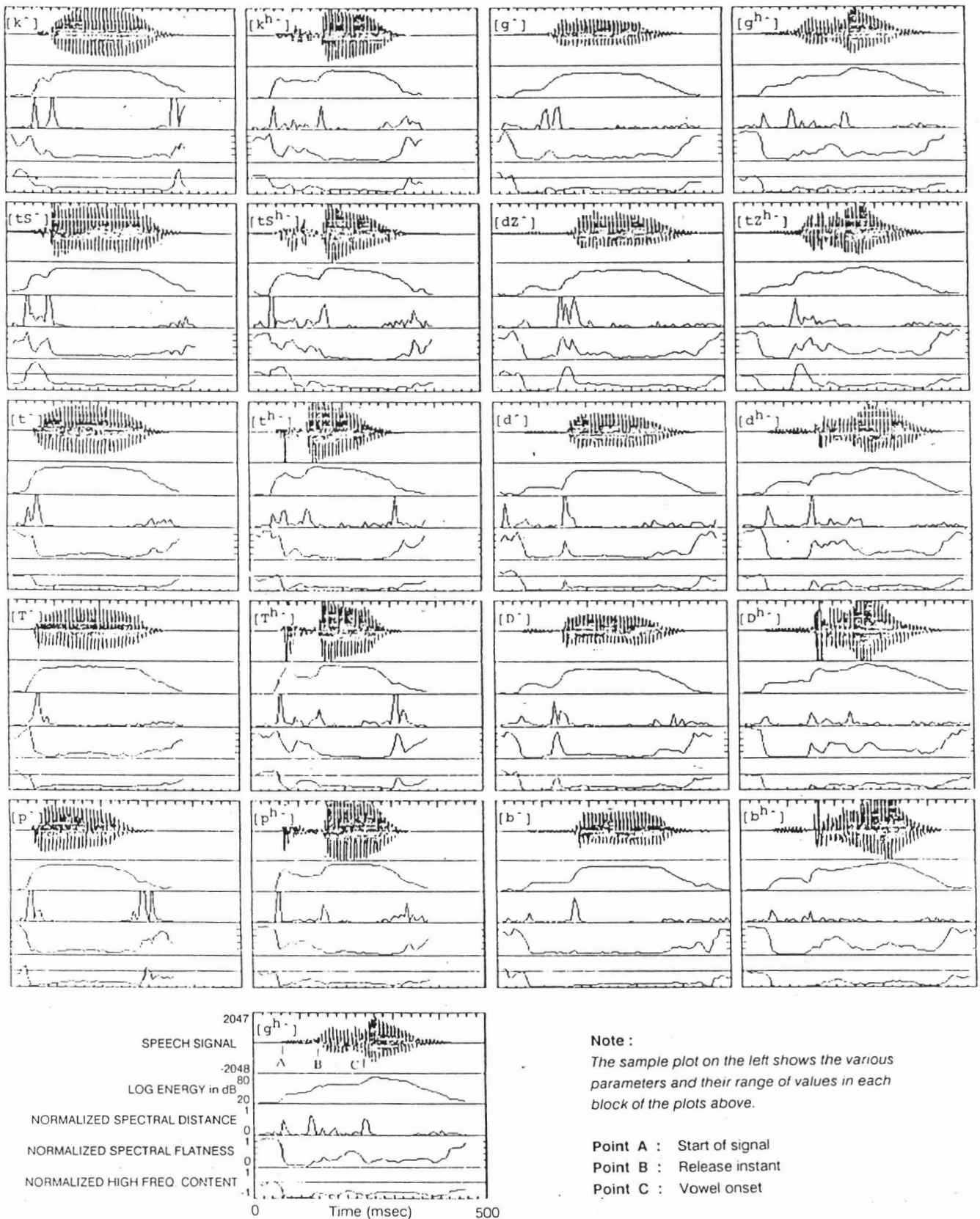
## **4.2 ISSUES IN DETERMINING SIGNIFICANT INSTANTS**

Fig. 4.1 shows some typical CV utterances of Hindi with the articulatory release instant and the vowel onset instant identified manually for one utterance. It can be seen that these instants demarcate distinct changes in the waveform plot. The figure also shows some of the gross parameters derived using block data processing and their trends for different utterances. The gross parameters shown in the figure are those identified at the end of this section. It is observed that though the change is distinct in the waveform plot, the parameters exhibit a smeared version due to inherent drawbacks of block data processing. Another point that may be observed is that, for utterances with very short consonantal segment as in [t<sup>h</sup>], [T<sup>h</sup>] and [p<sup>h</sup>], the smearing of change is such that we altogether miss the separation between the two events – start of the utterance and vowel onset.

The nature of the signal before and after the significant instants provide good clues to determine the important features that can lead to their detection.



**Fig. 4.1 : Plots of speech waveforms and gross parameters for utterances of Hindi stop consonants**



**Fig. 4.1 : Plots of speech waveforms and gross parameters for utterances of Hindi stop consonants**

The following subsections highlight these features for the detection of significant instants.

#### **4.2.1 Start of the utterance**

The region prior to the start of the utterance has usually no signal and is considered silence region. However, in practice, this region may exhibit very low amplitude variations due to background noise in the environment. The region immediately after the start of utterance exhibits different characteristics for different utterances. From the discussions in Chapter 3 on instants and regions it is seen that characteristic regions that occur after the start of the utterance are dependent on the manner of the consonant. The common trait irrespective of the manner is that there is an abrupt and significant rise in signal amplitude. One other feature is that there is a change from silence to a region with specific spectral characteristics exhibiting low frequency predominance in certain consonant manners and high frequency predominance in the other manners.

#### **4.2.2 The articulatory release instant**

In Chapter 3, we have seen that the articulatory release instant may co-occur with either the start of utterance or vowel onset instant or both in CV utterances of Hindi. The stop consonants (including affricates) have a strong epoch at the articulatory release instant because of the abrupt release of pressure buildup. The other consonants, namely, nasals, semivowels and fricatives, exhibit no such excitation epoch at the articulatory release instant. In some consonants, there is a distinct change in spectral characteristics because of change in vocal tract shape caused by the articulatory release. In all cases, the common feature is that an abrupt and distinct increase in the amplitude of the signal occurs at the articulatory release instant. But such changes in amplitude occur at other instants as well. It is necessary to qualify the abrupt amplitude increase with the adjoining context in the utterance. Therefore, detection of the articulatory release instant requires different kinds of processing for different cases.

### 4.2.3 The vowel onset instant

Onset of vowel in CV utterances is indicated by the onset of strong glottal vibrations with relatively open vocal tract. The speech signal before the vowel onset instant exhibits different characteristics for different consonant manners. Therefore, the common feature for any CV utterance is the region after the vowel onset instant. The speech signal after this instant has the vowel nature exhibiting periodicity and large amplitude. The significant clues for the vowel region are, presence of pitch epochs (which may be weak for some vowels), periodicity in the waveform, and good spectral structure due to the vocal tract resonances. Thus, the description for the vowel onset instant is that of onset of a characteristic region, or rather, change to a region with specific features. The spectral characteristics in the region after vowel onset are dependent on the vowel type of the utterance. Therefore to detect vowel onset, we need to look for the beginning of a region with the following features.

- a). Presence of pitch epochs (periodic excitation)
- b) Presence of spectral structure – region with low value of spectral flatness
- c) Periodicity in time domain waveform
- d) Large amplitude of glottal air volume flow
- e) High energy in the signal
- f) Gradual change in energy and
- g) Minimal variation in spectral features.

### 4.2.4 Summary

Based on the above discussions, the parameters listed in Table-4.1 can be used for characterizing the changes that occur on either side of the significant instants. The first parameter – normalized energy – gives indication of instants of distinct changes in waveform characteristics in the utterance at a gross level. The averaged zero crossings count is a coarse indication of the high frequency energy

relative to the low frequency energy in the signal. The normalized energy of differenced signal provides an indication of the trend in the high frequency energy. The spectral flatness measure is a first approximation to the gross spectral structure in the signal. Regions with distinct spectral structure as in vowel region exhibiting formants, are brought out as a low value in the spectral flatness measure. The first order LP coefficient is another approximation to the relative energy of the high and low frequency components in the signal. The spectral distance measure exhibits distinct peaks whenever there are distinct changes in spectral characteristics in the signal. The behavior of some of these parameters in different regions highlighting the distinct features of CV utterances can be clearly seen in Fig. 4.1. We shall now look at a signal processing method which can directly determine certain instants in the signal.

**Table—4.1 :** Parameters for characterizing significant instants

- |    |  |
|----|--|
| 1. | Normalized Energy  |
| 2. | Average zero crossing count                                      |
| 3. | Normalized Energy of differenced signals                         |
| 4. | Spectral flatness measure  |
| 5. | First order LP coefficient ( $-R(1)/R(0)$ )                      |
| 6. | Spectral distance between adjacent frames using Itakura measure. |

### 4.3 EPOCH DETECTION TO IDENTIFY SIGNIFICANT INSTANTS

Any signal may be viewed as the output of a source-filter model. The source in this framework is assumed to be idealized impulses and the spectral behavior in the signal is because of the filter's frequency shaping characteristics.

In other words, the signal is the result of response of the filter to impulse excitations. Speech signal may also be viewed in this source-filter context. Region of silence or background noise is characterized by signal with very low amplitude. Hence the source impulses (excitations or epochs) in this region are insignificant and weak. On the other hand, at the beginning of the utterance, there is presence of signal with significant amplitude and this is attributable to a strong source epoch. Hence the start of an utterance is indicated when a strong epoch is detected after a region with low or nil excitations. It may also be noted that speech signal is characterized by regions with different types of excitation and time varying filter characteristics. Hence other instants with strong epochs are also present in the signal after the start of an utterance.

Various approaches have been reported [Cheng 1989] [Ananthapadmanabha 1975] [Ananthapadmanabha 1979] for determining instants of significant excitation in speech signal with a view to obtain instants of excitation due to vocal cord vibrations during voiced speech. These have been carried out to determine either pitch periodicity or to obtain better estimates of the filter characteristics (as in pitch synchronous analysis and analysis of the signal during glottal closure duration).

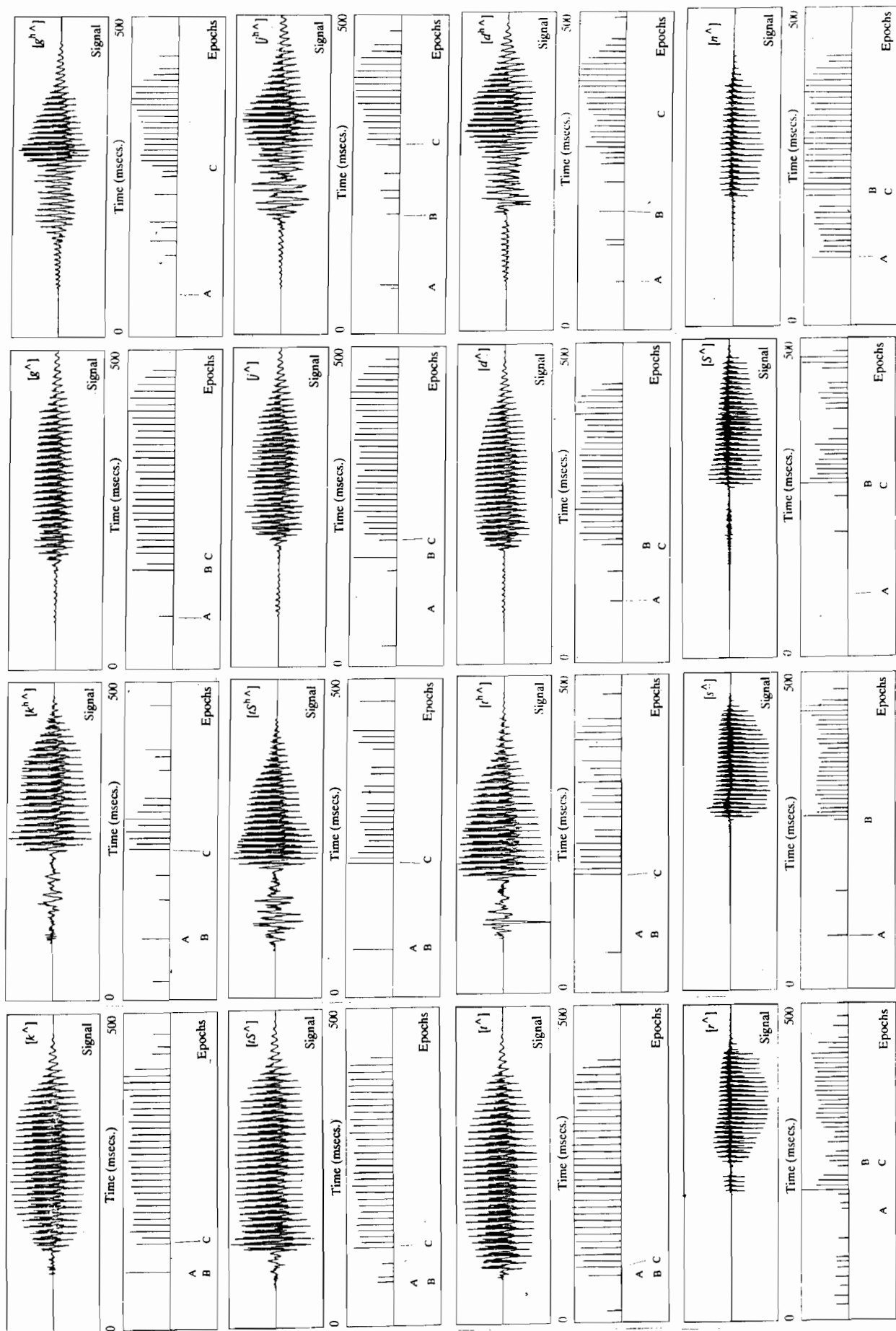
A recent approach by [Smits 1992] uses the minimum phase property of the signal and group delay processing methods to determine points of significant excitation in speech signal. Since speech is produced by a physical process, the signal after a significant excitation should exhibit decaying characteristics. This implies minimum phase property. If we consider a segment of speech signal taken arbitrarily, the instant of significant excitation may occur at any point in the signal. Such a signal would no longer be minimum phase unless the segment chosen is such that it begins with the excitation. Minimum phase property implies that among all possible signals with identical spectral magnitude response, the minimum phase signal would have the least average (minimum) group delay [Murthy 1991b]. The

approach in [Smits 1992] uses the average negative derivative of phase or group delay of the linear prediction residual of the signal to determine the instants of minimum phase (or rather the instants exhibiting significant minimum phase characteristics). The linear prediction residual is used as it is a good approximation to the excitation signal and because it minimizes the effects of the positioning of the window function with respect to the impulse response of the system.

Excitation in speech is due to a change in air volume flow. This change could occur due to abrupt pressure release as in the case of stop releases or due to turbulent flow as in fricative sounds and aspirated sounds or at glottis due to glottal vibrations during vowel production. During vowel production, maximum excitation in a pitch period is at the instant of greatest change in glottal air volume flow, which typically occurs at the instant of glottal closure. Other instants of change in glottal air volume flow are also points of excitation. Since one may view each instant in the signal as having an excitation due to the nature of the physical process, it is the strongest excitation (epoch) in the context of other nearby weak excitations which is significant and of interest. To determine the epochs, average group delay is computed at each instant in the signal. Negative zero crossings of this function are used to identify the instants of significant minimum phase characteristics or the epochs. A significant point about this epoch detection technique is that it yields information about instants in the signal with high resolution even though block data processing is used. The main disadvantage of this technique is that it is highly compute intensive. The average group delay needs to be computed at each sample point and this requires the computation of several DFTs.

Fig. 4.2 shows the significant epochs obtained using this method for a subset of the CV utterances. The points A,B and C marked in the figure identify the start of the utterance, articulatory release instant and vowel onset for these utterances. Sometimes the nature of processing results in spurious and irrelevant

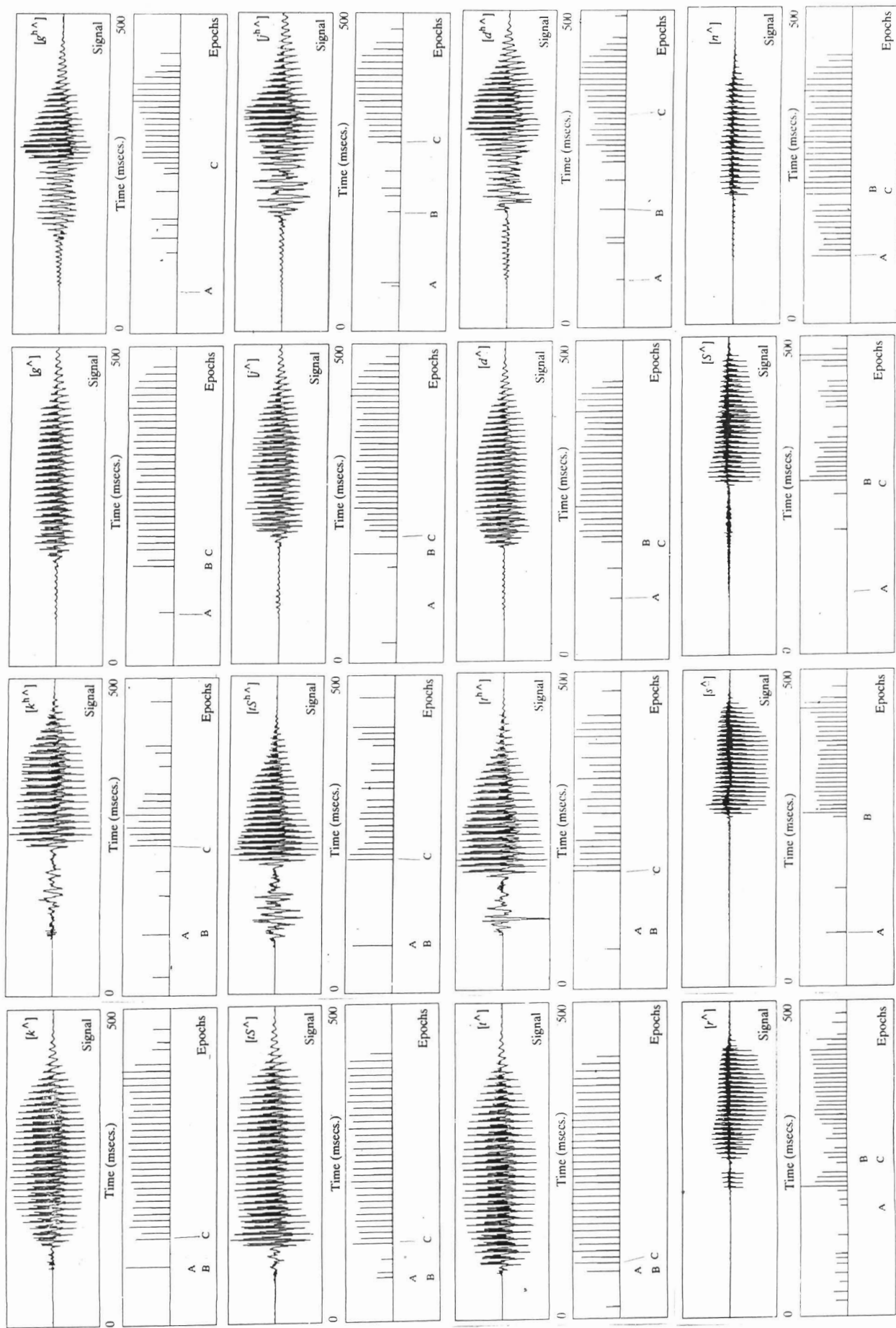




**Point A: Start of signal**

### Point B: Articulatory release instant

### Point C: Vowel onset instant



Point A: Start of signal  
 Point B: Articulatory release instant  
 Point C: Vowel onset instant

Fig. 4.2 : Speech waveform and Epochs (instants of significant excitation) derived using the phase derivative method for some CV utterances

epochs which need to be filtered using context and other parameters. The epochs determined using this method aid in the manual identification of the significant instants and have been used as such extensively, in the present study. The following discussion highlights our observations on the performance of this technique in identifying the three significant instants of interest in our study.

It was observed that most of the problems in epoch detection occur in the consonantal part of the utterance. The pitch epochs in the vowel region are usually identified except in rare cases as seen in Fig. 4.2 for  $[k^h\wedge]$   $[tS^h\wedge]$  and  $[t^h\wedge]$ . Epochs at vowel onset are missed occasionally. Random epochs are identified in the middle of the aspiration and frication regions occasionally. Noise in silence regions is identified frequently as epochs. The start of the utterance is not identified as an epoch in the case of voiced stop and fricative consonants at times. Some of these observations can be seen in the plots of Fig. 4.2. These observations show that post processing is called for before we can use the epochs as corresponding to significant instants.

#### **4.4 DETERMINATION OF INSTANTS BASED ON PARAMETRIC CHANGE**

We have seen from Fig. 4.1 that gross parameters derived using block data processing do not yield the significant instants with good resolution. But, they suffice for our requirement of identifying regions for detailed analysis. Since gross parameters are usually some form of averaged values, they mask details and bring out general features. Hence, they generally show similar trends for any CV context. Methods based on the use of gross parameters for end point detection, voiced/unvoiced/ silence detection and segmentation have been used both in isolated word and continuous speech recognition [Rabiner 1975] [Atal 1976]. These methods generally use somewhat arbitrary thresholds on the parameters and suitable logic modules for validation. An alternative proposed here is to make use of a multilayer perceptron artificial neural network to perform the

identification of significant instants in CV utterances. To realize detection of instants using this approach, the ANN is trained using parameters from either side of the significant instants. The training procedure requires manually identified instants in the training data. Though this calls for identifying instants of change during training, once the network learns the context of change, it is possible to feed the parameters from any region during actual usage. The network would respond with an active output only on the occurrence of significant change in the parameters representing the significant instant. A network with two hidden layers is used as it is sufficient to learn arbitrary decision surfaces.

#### **4.5 DISCUSSION OF RESULTS**

The gross parameters listed in Table 4-1 were used as inputs to the network. The first five parameters are derived separately from frames on either side of the significant instant. The sixth parameter is the distance measure across the significant instants. In all, eleven parameters were used as input for the network. The outputs of the network correspond to the significant instants. One output is identified with each significant instant. Therefore there are three outputs in the network. Two hidden layers are used with fifteen and ten nodes in the first and second hidden layer respectively. The network was trained using parameters from either side of the manually identified instants of the CV utterance. Since some of the instants occur together as discussed in Chapter 3, the neural network training was carried out taking this also into account. The first study was conducted by testing the trained network with parameters from manually identified boundaries of the signal. Since more than one output in the network can be active, it is not possible to use the output with maximum activation level as the indication for the identified instant. Each of the network outputs is compared against a threshold. Outputs with activation levels higher than the threshold were taken as indication that the corresponding instants were identified. Different threshold values were tried ranging from 0.3 to 0.7. The performance of the network with threshold values

set to 0.3, 0.5 and 0.7 are shown in Table-4.2. The table shows the performance of the network in terms of the correctly identified and wrongly identified instants as percentages. The percentage of missed identifications are not included as they can be read as the complement of correctly identified percentages. It is seen that with lowered detection threshold the recognition performance does increase, but consequently the wrong identification percentage also increases. Conversely, with the detection threshold set high, the percentage of correct identification decreases with an attendant decrease in the wrong identification.

**Table-4.2:** Verification of performance of identification of significant instants in percentage values

	DETECTION THRESHOLD	CORRECT IDENTIFICATION	WRONG IDENTIFICATION
Start of utterance	0.3	91.7	17.3
	0.5	87.4	12.6
	0.7	76.2	7.5
Articulatory release instant	0.3	98.0	22.2
	0.5	94.8	17.2
	0.7	89.3	8.3
Vowel onset instant	0.3	94.5	12.8
	0.5	91.6	8.1
	0.7	82.2	4.4
OVERALL	0.5	91.26	12.8

A threshold value of 0.5 seems to be a reasonable compromise between the correct and wrong identifications. It is seen from the table that the correct identification of start of utterance is poorer. It was noticed that many of the missed instants occurred for the case of voiced stops and fricatives. This is attributed to the low amplitude of the signal (and consequently SNR) which causes poor

parametric representation across the instant. Maximal errors in wrong identification occurs for the vowel onset instant. The cause for this is attributed to the varying nature of the vowel onset instant for the different consonant manners. The overall performance shows fairly good performance for the important significant instant namely, the articulatory release instant. This instant needs to be determined reliably since the analysis for the place of articulation is dependent on this. It should be noted that the results are reported for tests carried out with data from manually identified instants for verifying the performance of the trained neural network for this purpose. Ideally, after training, the network should be usable by hypothesizing every instant in the signal as a significant instant and verifying the hypothesis by observing the outputs of the network.

With these significant instants as reference points to identify regions in a CV utterance, it is possible to carry out pertinent detailed analysis relevant to the regions. Spectral characteristics of the signal in the region after vowel onset can be used to **determine** the vowel type of the CV utterance. Vowel type determination using this approach is discussed in Chapter 5. Gross characteristics of the signal in the region before vowel onset can be used as cues for classification of the consonant in terms of the manner of articulation. This forms the topic of discussion in Chapter 6. Changes in spectral characteristics in the region immediately following articulatory release instant, provide the cues to determine the place of articulation. Use of this key feature is discussed in Chapter 7. Combining the results of all the three stages, it is possible to recognize CV utterances.

## **STUDIES ON RECOGNITION OF VOWELS IN CV UTTERANCES**

The objective of the studies reported in this chapter is to explore methods for determining ~~recognize~~ the vowel class from speech data in the vowel region of a CV utterance. The steady region in a CV utterance after the vowel onset can be considered as the vowel part. Irrespective of the consonant context, utterances with the same vowel ending show similar characteristics in the vowel part, but for the initial transitions. Therefore, the region in the vowel part after a short (30 msec) interval from the vowel onset (to avoid the transition region) is used for analysis to determine the class of the vowel. As indicated in Chapter 3, there are different types of vowels in Hindi of which two are diphthongs and three others are distinguished mainly by their length. From articulatory description of the vowels of Hindi, as given in Table-3.1 (Chapter 3), as well as from vowel quality considerations, it is seen that the Hindi vowels, in general, can be categorized into five classes. Therefore in our studies, only the CV utterances ending with vowels belonging to the five classes – [I], [e], [^], [o] and [U] – have been considered. In the context of nasal utterances, the vowels and diphthongs exhibit nasalization. Since nasalized vowels form a different group, they have not been included in our studies. The next two sections describe the issues involved in vowel discrimination and the parameters chosen for vowel discrimination. Section 3 describes as to how a neural network classifier can be used for vowel discrimination and also provides details of data used for training and testing. The last section discusses in detail the results of the studies conducted on the recognition of vowels in CV utterances.

## 5.1 ISSUES IN VOWEL DISCRIMINATION

Vowels are discriminated by their quality, from the human perception point of view, and more specifically by the shape of the vocal tract during its production. From a phonetician's point of view [Fant 1970], [Ladefoged 1975], [Fry 1979], [Catford 1988], it is customary to give a gross description of the vocal tract configuration (shape), in terms of the position of the tongue hump, the extent of closure due to the tongue hump and the shape formed by lips. This description usually suffices to unambiguously classify the different vowels. This description has fuzzy demarcations which reflect the natural variations that occur in different instances of the same sound. A point of interest in this form of description of the vocal tract configurations of different vowels is that it implies three independent dimensions of differences. The vocal tract configurations for the vowels of Hindi are as given below:

Position of tongue hump	: front, central and back;
Closure due to tongue hump (described as height)	: low, mid and high;
Shape of lips	: rounded and unrounded.

Though independent gradations in these dimensions imply a large possible set of vowels, the actual vowel categories in a language are limited. Though vocal tract shape is the defining feature for vowel discrimination, what is realized in the speech signal is acoustic filtering due to the shape of the vocal tract. As discussed in Chapter 3, the vowels are normally described in terms of the spectral characteristics that manifest in the speech signal because of this filtering.

### 5.1.1 Manifestation of speaker characteristics

An important aspect that needs to be considered when attempting discrimination between the vowels is that of handling speaker independence. The vowel quality is both due to the vocal tract shape and the physical dimensions of the tract. The physical dimensions of the tract vary from speaker to speaker. Thus,



the vowel quality has imbibed in it the speaker's characteristics as well. Different vowels may be viewed as arising out of changes in the shape of a neutral vocal tract. This viewpoint suggests that by considering the range of variations in discerned tract dimensions for the different vowels, one can achieve vowel discrimination in a speaker independent manner.

### **5.1.2 Cues for vowel discrimination**

A standard approach to vocal tract shape determination is by modeling the speech production process as a source exciting a filter. The filter's response is equated to the acoustic resonances of the vocal tract. The formants are the gross spectral peaks that occur due to the resonances. The resonances thus obtained can be considered as a representation of the vocal tract shape, since the shape and dimensions dictate the resonances. Several acoustic phonetic studies [Fant 1970] [Catford 1988] have indicated that the values of the first two formants **F1** and **F2** provide reasonable discrimination among vowels. Fig. 5.1 shows the positions of the Hindi vowels in the **F1–F2** plane for a male speaker. The figure also shows the averaged positions for the same vowels. But there is appreciable variation when considering across speakers which results in discrimination errors. An improvement in discrimination may be achieved by considering the difference of **F2** and **F1** as another parameter [Ladefoged 1975]. Errors in vowel discrimination, when considering several speakers, based on **F1** and **F2** have led some researchers to question the basis for the use of **F1** and **F2** alone [Klein 1970]. Some studies by [Fant 1970], [Carlson 1979], [Kuhn 1975] have indicated that an important perceptual cue to the vowel quality is the front cavity resonance of the vocal tract and this corresponds to the pseudo second formant. This has led to the consideration of pseudo second formant rather than **F2** for vowel discrimination.

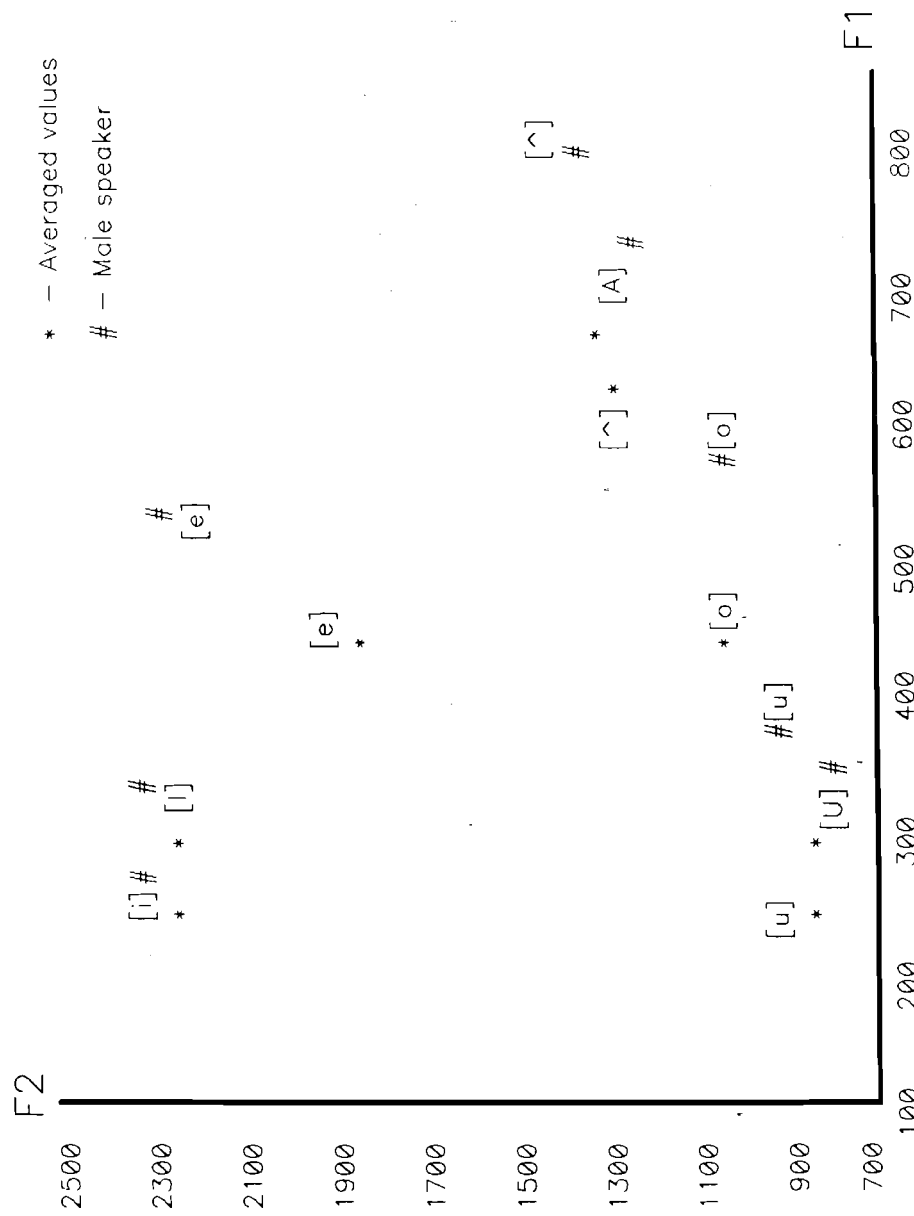


Fig. 5.1 : Hindi vowels in F1-F2 plane

## 5.2 CHOICE OF PARAMETERS FOR VOWEL CLASSIFICATION

Several techniques have been proposed for ~~the~~ determination of formants using block data analysis [Rabiner 1978b]. These approaches use methods to determine peaks in the smoothed spectral envelope of the signal. Linear prediction based spectral smoothing and ~~cepstral~~ smoothing methods have been ~~used~~. Time domain methods using zero crossings count of ~~bandpass~~ filtered signals have also been proposed [Niederjohn 1985]. The formants are picked from the spectral peaks and validated based on a decision logic which uses constraints in the range of expected formant values as well the differences with respect to the formant values in neighboring frames of the signal.

Pitch synchronous analysis has been used to obtain better estimate of formants by using the signal during glottal closure [Parthasarathy 1987]. But this requires determining pitch epochs (glottal closure instants) prior to analysis. Good results have been obtained using non-model based approaches applying group delay domain processing methods [Murthy 1991a]. Time-varying analysis methods have been proposed to improve performance as against short-time invariant assumptions [Nathan 1991]. Perceptual linear prediction has been proposed where the spectral features are related to the pseudo second formant [Hermansky 1986].

### 5.2.1 Gross spectral features for vowel discrimination

In the context of discrimination among a limited number of vowels as in the present case, it is simpler and more appropriate to use the gross spectral structure rather than the formant values. ~~The~~ variability in the values of the formants is conveniently masked in the gross spectral structure. An added benefit is the absence of errors due to wrong estimation of formant values. Linear prediction analysis with about 10 to 14 coefficients, for the vowel region of speech signal sampled at 10 kHz provides a good all-pole fit. Cepstral coefficients derived from the linear prediction coefficients are even better as they represent a better spectral

fit. Weighted cepstral coefficients are used to **compensate** for the progressive decay in the values of the higher numbered coefficients [Yegnanarayana 1979]. A more direct cue to vowel discrimination is the vocal tract shape as **discussed** earlier. Hence area coefficients describing the vocal tract shape are studied in detail.

Acoustic-phonetic descriptions use the vocal tract shape to describe the vowel quality in the speech signal. This implies that the shape feature for different vowels is speaker independent. Several researchers [Wakita 1973], [Atal 1971] have studied the use of linear prediction analysis to directly estimate the discrete area function representation of the acoustic-tube model from the speech signal. The discrete area function representation commonly termed as the area function, gives the areas of the different sections (area coefficients) of the acoustic-tube model. [Wakita 1975] has shown that good correspondence between the estimated area coefficients and the gross vocal tract shape for different vowels is obtainable, if the speech signal is properly pre-emphasized and proper boundary conditions of the acoustic tube model are chosen. In this formulation, the acoustic tube is modeled as concatenation of discrete sections of tubes of equal length but differing areas. If  $f_s$  is the sampling rate of the signal, then the sampling period  $T$  is related to the section length  $l$  by the following equation.

$$T = \frac{2l}{c}, \quad (5.1)$$

where  $c$  is the velocity of sound in the acoustic tube. Multiplying both numerator and denominator by the number of sections,  $M$ , yields

$$T = \frac{2Ml}{Mc} \quad (5.2)$$

Now,  $Ml$  corresponds to the overall length of the acoustic tube,  $L$ . Replacing  $Ml$  by  $L$  and  $1/T$  by  $f_s$  in the above equation and rearranging the terms results in the following equation for the number of sections  $M$  :

$$M = \frac{2f_s L}{c} \quad (5.3)$$

Studies by [Fant 1970] indicate that the average length of the human vocal tract is about 17.5 cms. Taking the value of  $c$  to be 34 cm/sec. (again from the same source), the number of discrete sections for the acoustic tube model at 10 kHz sampling rate is ten. Thus, ten sections are deemed sufficient for 10 kHz sampling rate. In this formulation, the number of sections is directly related to the order of linear prediction. It should be noted that to obtain an increased resolution of the area coefficients, in terms of the number of sections, the sampling rate should be increased correspondingly.

The adjacent section area ratios are inversely related to the acoustic impedance ratios of the two sections which are in turn related to the reflection coefficients of the acoustic tube model. The relationship between the area ratio and acoustic reflection coefficient is given below.

$$\frac{A_m}{A_m + 1} = \frac{(1 - \mu_m)}{(1 + \mu_m)} \quad (5.4)$$

where  $A_m$  is the area of  $m$  th section ( $m=1$  corresponds to the lips end and  $m=M$  corresponds to the glottis end of the acoustic tube), and  $\mu_m$  is the acoustic reflection coefficient at the boundary of  $m$  and  $m+1$  sections.

As per the formulation, the  $\mu_m$ s directly correspond to the reflection coefficients  $k_m$ s obtained through linear prediction analysis. Consequently the area ratios are related to the reflection coefficients  $k_m$ s obtained from linear prediction analysis by the equation

$$\frac{A_m}{A_m + 1} = \frac{(1 - k_m)}{(1 + k_m)} \quad (5.5)$$

with the range of  $k_m$ s being  $-1.0$  to  $+1.0$ .

Thus it is possible to obtain the area ratios of adjacent sections of the acoustic tube from reflection coefficients obtained by linear prediction analysis.

The actual area is not important and relevant but, by making suitable assumptions about the area at the boundary of the acoustic tube, it is possible to build up the area coefficients from the area ratios of the adjacent tube sections using the reflection coefficients. Physical constraints imply that the area coefficients should be nonzero and positive. It is seen from equation (5.5) that nonzero, positive areas **can** be realized, provided the magnitudes of the reflection coefficients are less than unity. Linear prediction analysis using the autocorrelation method yields reflection coefficients which satisfy this condition [Markel 1974]. One problem that arises when using these equations is that for values of  $k_m$  with magnitude close to unity, the area ratio indicates a very large change in area. The dynamic range of the resultant area coefficients becomes very large. The vocal tract does not exhibit very large area changes. The errors in the acoustic tube model arise mainly due to inherent errors in the assumptions. The main problem of this large dynamic range of the area coefficients is the resultant variability in the area coefficients (areas of different sections of the acoustic tube model) even for similar vowels. Several approaches have been used to reduce the dynamic range of the area ratio [Viswanathan 1975] [Markel 1974] in speech coding applications.

Since magnitudes of  $k_m$ s are close to 1 for roots close to the unit circle in the z-transform plane, the magnitudes of  $k_m$  can be reduced by pushing all the roots (poles) inwards in the unit circle which causes increased damping. This manipulation can be realized by multiplying the linear prediction coefficients  $a_i$  by a factor  $r^i$  where  $r$  is a positive quantity less than 1. A first approximation to this can also be realized by adding a small constant to the autocorrelation coefficient  $R(0)$  prior to determining the linear prediction coefficients. The latter approach has the effect of damping the poles closer to the unit circle more than those close to the origin. It was found that both the approaches result in distinct reduction in the dynamic range of the area coefficients. But, the reduction in variability is not uniform. For the area coefficients close to the lips end, the reduction in variability is less whereas, close to the glottis end, the variability is considerably reduced.

An **effective** method for **reducing** the dynamic range of the area ratios is by considering a variation of the expression for the area ratio. Instead of defining the area ratio as  $(1-k_m)/(1+k_m)$  one can compute  $(2-k_m)/(2+k_m)$  [Markel 1974]. The effect of this change is that the range of area ratios is  $(0.3 < \text{ratio} < 3)$  rather than  $(0 < \text{ratio} < \infty)$ . Since this operation is merely one of range compression, the **shape** feature is not significantly altered. Area coefficients derived using this expression have been used in our study to see the effectiveness of the area coefficients to overcome speaker dependence. It is necessary to carry out pre emphasis of the speech signal prior to linear prediction analysis, to correct for the combined effect of lip radiation and glottal source characteristics. An adaptive pre-emphasis as suggested by [Viswanathan 1975] has been used. Pre-emphasis is carried out by using a first order inversefilter with the transfer function  $(1 - az^{-1})$  where 'a' is given by  $R(1)/R(0)$ .  $R(0)$  and  $R(1)$  are the autocorrelation coefficients with zero and one sample lag respectively.

### 5.2.2 Parameters for vowel discrimination used in this study

In our studies, modified form of area coefficients as well as weighted cepstral coefficients derived from linear prediction analysis have been used to parametrize the spectral characteristics of the vowel region. The objective was to use area coefficients for classification as they are closer to the articulatory features. The weighted cepstral coefficients were used for comparison. The analysis region is a short segment of the speech signal in the vowel region. Typically this region has large amplitude signal and consequently a high signal to noise ratio. Therefore it is possible to extract spectral parameters reliably. In our study, one frame of speech signal from about 30 ms after the vowel onset instant is used. The frame size for analysis was chosen to be 10 ms. To have better representation of the spectral characteristics of the signal in the vowel region, a near pitch synchronous analysis was realized using the short time energy trend in this region of the speech signal. The frame is chosen such that short time energy has a peak in the center of the frame. The signal is windowed using a Hamming window. A signal

dependent preemphasis is performed as indicated in the preceding subsection. The cepstral coefficients are obtained from the LP coefficients recursively. The area coefficients are obtained using the reflection coefficients with the glottal end of the tube as reference.

The plots of the linear prediction **spectrum** and the area coefficients are shown in the first and second columns of Fig. 5.2. These are determined for the five vowels of Hindi, spoken by a male speaker for a large number of repetitions. The figure illustrates the nature of **variability** in these parameters for several repetitions of these utterances by the same speaker. Fig. 5.3 illustrates the characteristics for the case of a female speaker.

### 5.3 VOWEL CLASSIFICATION USING A NEURAL NETWORK CLASSIFIER

It is seen from the illustrations in the previous section that there is significant variability in the spectral features of the speech signal in the vowel region even for the same speaker. Therefore, for the recognition of the vowel category, we require a classifier which can capture these features despite their variability and discriminates the vowel classes well. We have used a neural network classifier for this purpose. A neural network can learn the complex decision **surfaces** for classification. A network with two hidden layers can realize any arbitrary decision surface [Lippman 1987] [Hush 1993]. In our studies on vowel classification, a feed forward neural network with two hidden layers has been used. The network is trained using back propagation algorithm. The network parameters were determined by trial and error. No attempt has been made to optimize the network. The number of inputs is determined by the dimensions of the parametric representation. The number of outputs is determined by the number of classes. As indicated in the preceding sections, we have used two different **representations** of the vowel region in our studies. The input dimension for both these cases is the same. Since a 10th order LP analysis has been used, the number of cepstral



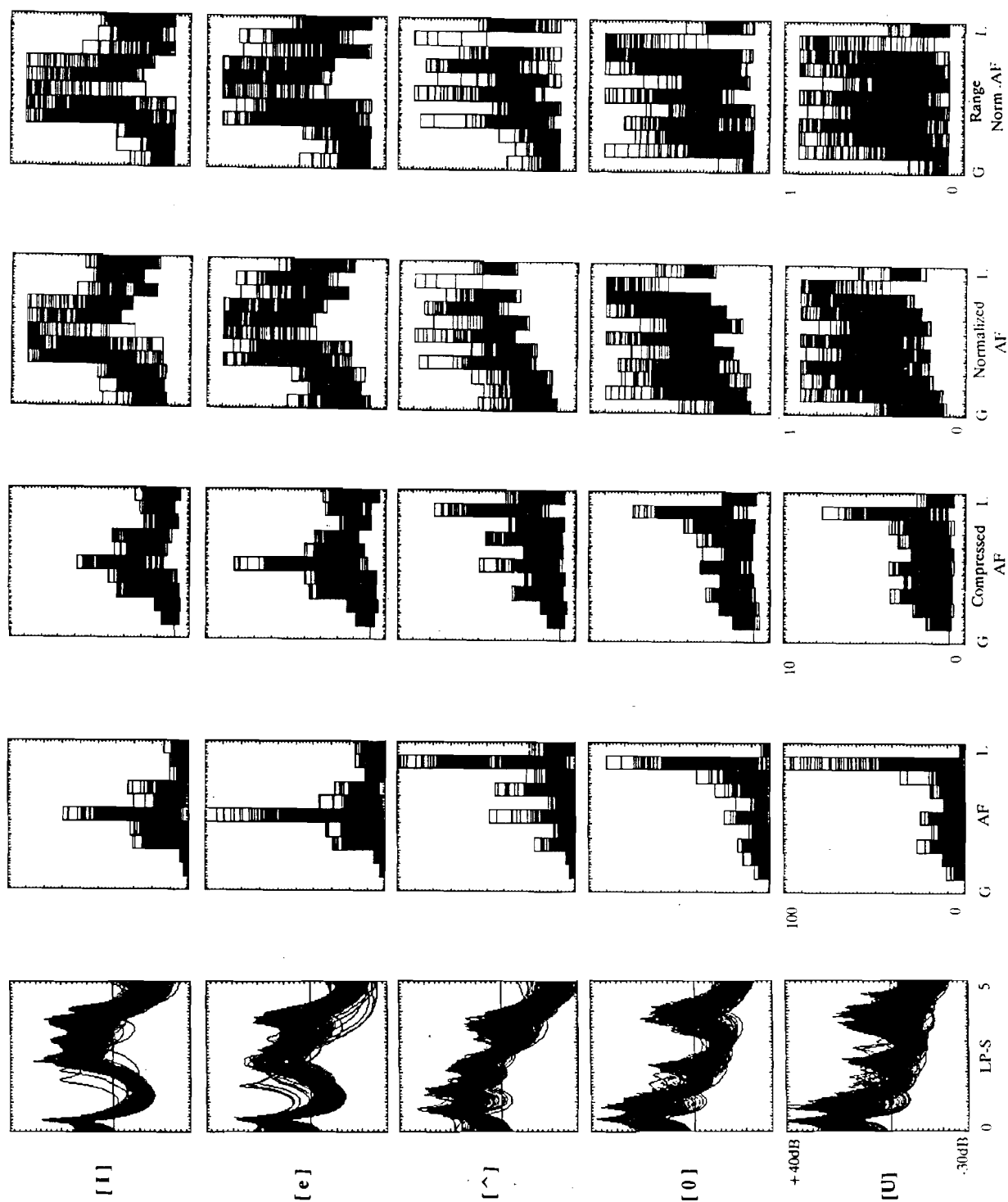


Fig. 5.2: LP spectrum and different normalization of the area function for many repetitions of the different vowels by a Male speaker

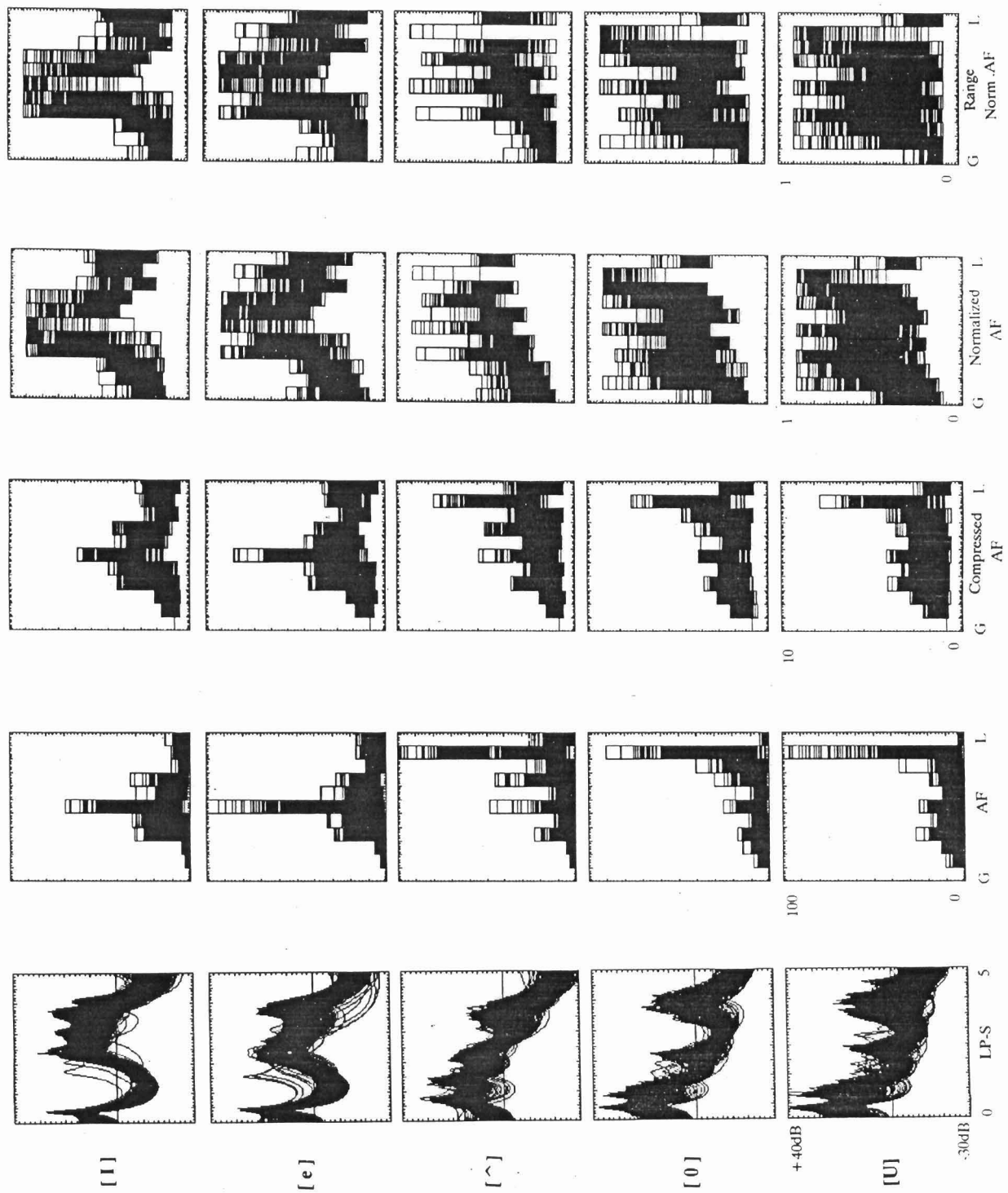


Fig. 5.2: LP spectrum and different normalization of the area function for many repetitions of the different vowels by a Male speaker

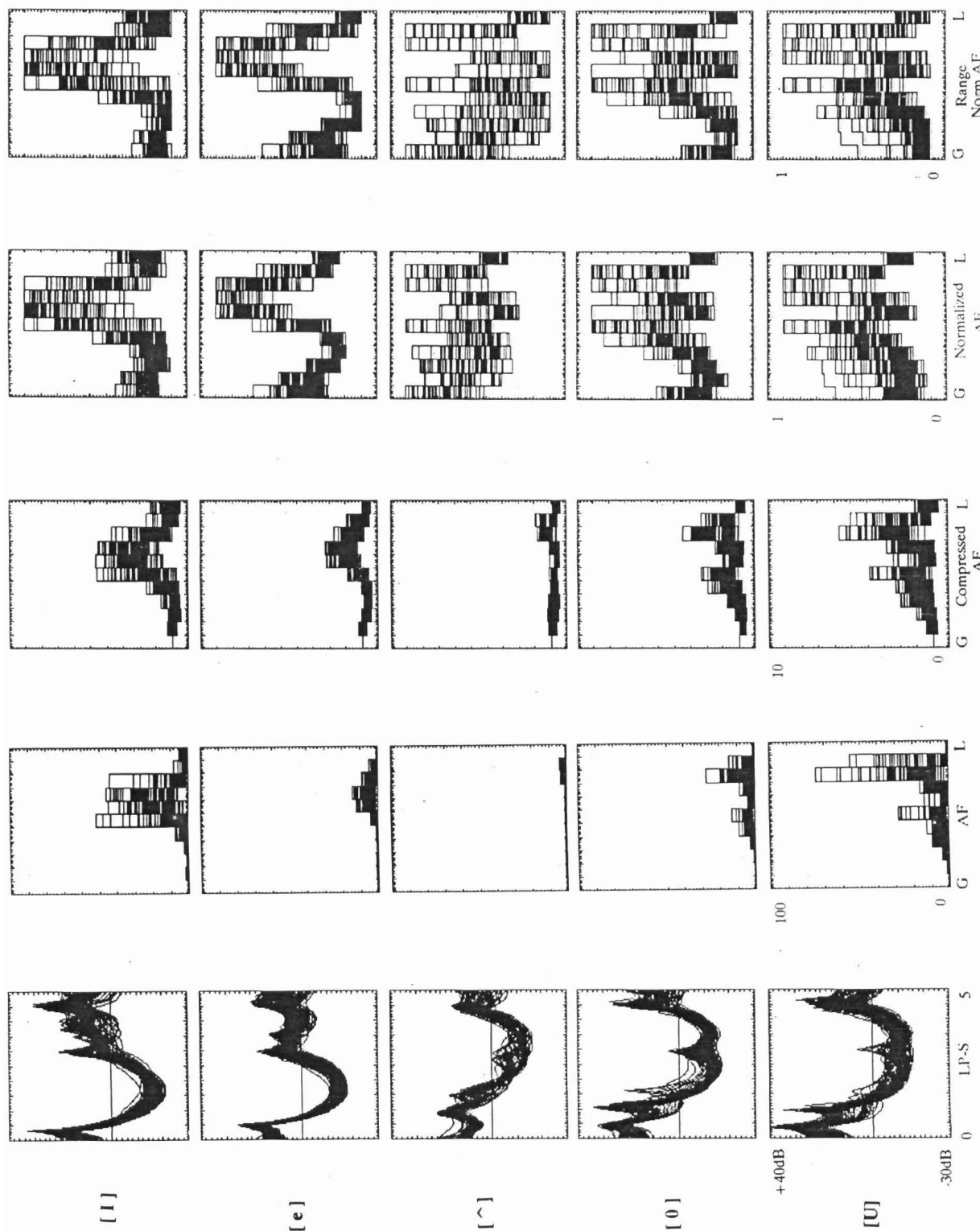


Fig. 5.3: LP spectrum and different normalization of the area function for many repetitions of the different vowels by a Female speaker

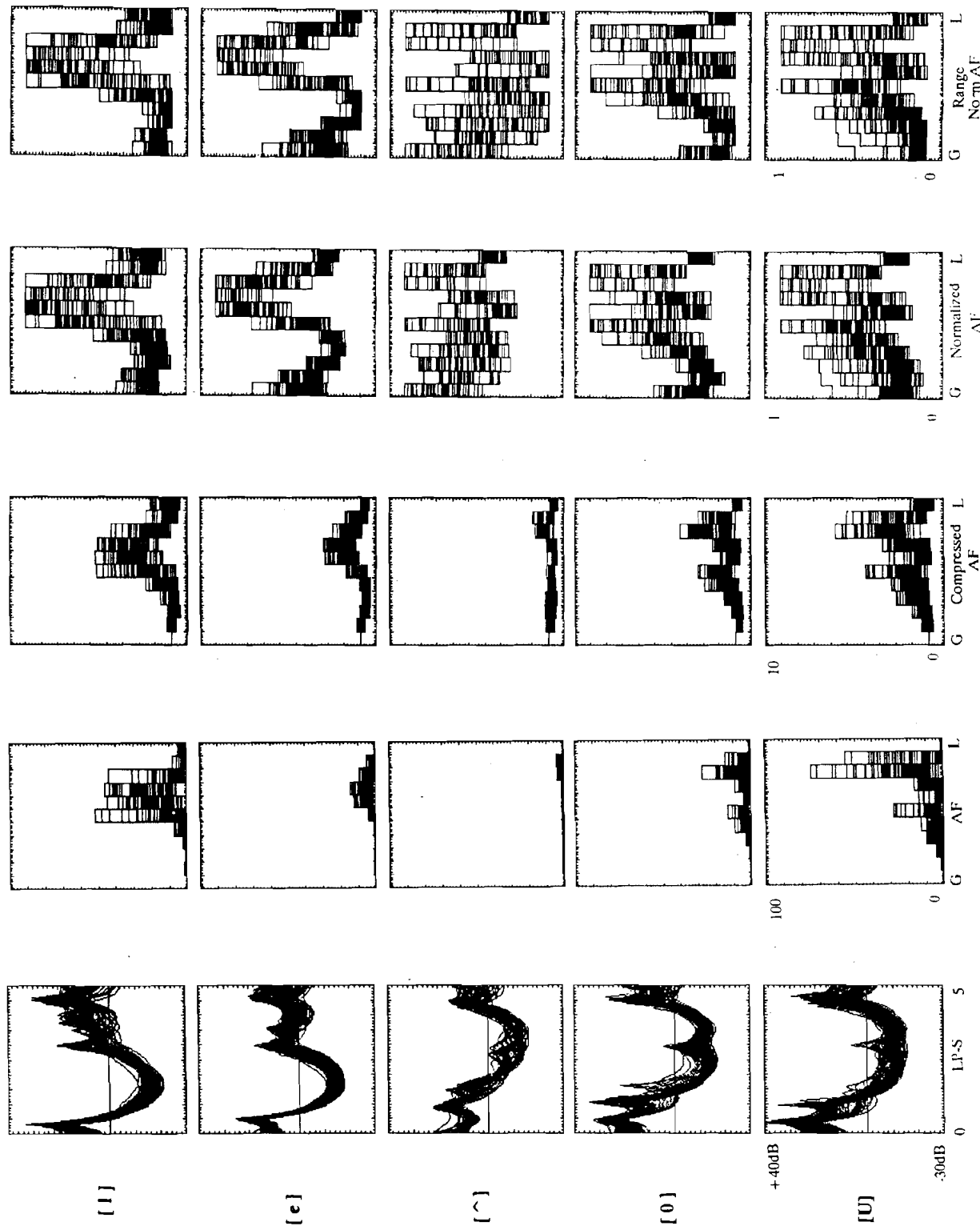


Fig. 5.3: LP spectrum and different normalization of the area function for many repetitions of the different vowels by a Female speaker

coefficients and area coefficients are ten each. Therefore, the number of inputs to the network is ten in each case. Since the number of vowel classes in our study is five, the number of output nodes is also chosen to be five with one output corresponding to each vowel class. The free variable in the network structure is the number of nodes in the hidden layers. The first hidden layer is chosen to have 30 nodes and the second hidden layer is chosen to have 20 nodes. These choices were guided by the heuristics and some preliminary studies. The network was initialized using random interconnection weights. The training procedure is the standard back propagation algorithm with the learning rate and momentum factors set suitably.

### 5.3.1 Data preparation for the classifier

In neural network classifiers, it is preferable to normalize the input data so that the input values are neither close to zero nor too large. Cepstral coefficients other than  $C_0$  are inherently normalized. Hence these values can be used directly as inputs to the neural network. Area coefficients on the other hand, even after compression, exhibit large values in many cases (as shown in Fig. 5.2 – 5.3). It is necessary to reduce the range of variation so that it falls within unity. This can be performed in several ways. One approach is to normalize the maximum area in each area coefficients to unity and scale the others correspondingly. This has the effect of deemphasizing the variations in small areas and seems to give importance to the large areas (see column 4 of Fig. 5.2). A second approach to normalization is to scale the area coefficients by a fixed factor. The choice of this second method is driven mainly by the consideration that the area coefficients is built up with the glottis end as the reference. Generally for any speaker, vocal tract area does not change significantly at the glottis end in the context of different vowels. Therefore, any normalization should preferably retain this reference. This approach is equivalent to using the area coefficients directly. A third approach to normalization is possible. In this approach, the effective range of variation of the area values are scaled such that any area coefficients has the same minimum

and maximum values. The choice of the third approach ~~is~~ inspired by the notion that in the area coefficients, the more important feature ~~is~~ its gross shape. The gross shape is determined by the position of maxima and minima in the area coefficients. Therefore normalizing the area coefficients with, ~~its~~ maximum and minimum values as reference would result in comparable area coefficients shape. The latter two forms of normalization have been used throughout in our studies. Columns 3 and 5 of Fig. 5.2 illustrate the effect of these methods of normalization when compared to using actual area coefficientss. The plots are based on many utterances of the vowels for a single speaker situation.

### 5.3.2 Recognition studies

Different networks were used for the three different parametric representations but all of them used the same structure. As the context is that of isolated utterances, it was assumed that all vowels occur with the same probability. The network was trained with the same number of patterns for each class. The training data set used in the study for one speaker consists of **140** patterns. The CV utterances considered for a vowel include all consonant contexts. Twenty ~~eight~~ of these patterns from each class were used for training and the remaining patterns were used to test the network classification performance. During the training procedure, the target output for each input pattern is known apriori. This was used in the back propagation algorithm to adjust the interconnection weights. Typically the learning process is iterated many times with the trained patterns. In our studies, it was found that beyond about 600 iterations of the learning procedure, the error term decreased at a very slow rate. Nevertheless, 1000 iterations were carried out.

## 5.4 DISCUSSION OF RESULTS

The testing of the network was carried out by recording the activation level of each output node for the different input patterns of the test data set. As

a first level discrimination, the output with the highest activation level was taken as representing the recognized class. Table-5.1 shows the confusion matrices for the three parametric representations using this approach.

The confusion matrix is organized such that the adjacent vowels in the table are closer in terms of quality and shape, and correspondingly closer in terms of the features as well. The table illustrates that confusion in discrimination occurs mostly with the neighboring vowel. It is seen from Table-5.1 that the weighted cepstral coefficients show very high recognition score. The area coefficients shows a degraded performance mainly because of the variability in the area coefficients. The degradation in performance of the range normalized area coefficients is because of the poor score for vowel [U]. These results are for the case of data obtained from a single speaker.

To see if similar performance results can be obtained for any speaker data, the study was conducted with another male speaker's voice and a female speaker's voice. Tables-5.2 and 5.3 show the confusion matrices for identical network structures which are trained and tested for another male speaker and female speaker's data using the same three parametric representations. Comparing these results with that in Table-5.1, it is seen that the performance results are nearly the same for the cepstral coefficients and the compressed area coefficients. The range normalized area coefficients shows better score compared to the compressed area coefficients for these cases.

The next study is to check whether the network trained for one speaker's data performs discrimination for another speaker's data as well. Table-5.4 lists the performance realized when testing the patterns of one speaker using the network trained with the patterns of another speaker. The results show degraded performance for all the three parametric representations. Another study was carried out to see if the network can learn the features better if both speakers' data are

**Table-5.1 : Performance of vowel discrimination - Speaker A**  
**(Confusion matrix in percentage values )**

**Male Voice**

a) Cepstral coefficients (CEP)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	84.28	12.86	0.00	0.00	2.86
[e]	0.00	100.00	0.00	0.00	0.00
[ ^ ]	0.00	0.00	100.00	0.00	0.00
[o]	0.00	3.57	2.14	92.86	1.43
[U]	0.00	0.71	0.00	5.00	94.29
				OVERALL	94.28

b) Area function - compressed (AFC)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	84.28	15.72	0.00	0.00	0.00
[e]	0.70	99.30	0.00	0.00	0.00
[ ^ ]	0.00	0.70	70.72	14.29	14.29
[o]	0.00	0.70	4.29	87.14	7.87
[u]	2.15	1.42	0.00	10.00	86.43
				OVERALL	85.57

c) Area function - compressed and range normalized (AFCR)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	88.57	11.43	0.00	0.00	0.00
[e]	0.70	99.30	0.00	0.00	0.00
[ ^ ]	0.00	0.00	81.43	15.72	2.85
[o]	0.00	2.15	3.57	82.14	12.14
[U]	0.00	3.57	2.15	48.57	45.71
				OVERALL	79.43



**Table-5.2** : Performance of vowel discrimination - Speaker B  
(Confusion matrix in percentage values)

**Male Voice**

a) Cepstral coefficients (CEP)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	98.14	0.93	0.93	0.00	0.00
[e]	4.63	95.37	0.00	0.00	0.00
[ ^ ]	0.00	2.78	97.22	0.00	0.00
[o]	0.00	0.00	2.78	92.37	1.85
[U]	0.00	0.00	0.00	6.48	93.52
				OVERALL	95.93

b) Area function - compressed (AFC)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	87.96	9.26	2.78	0.00	0.00
[e]	5.55	91.67	2.78	0.00	0.00
[ ^ ]	0.00	6.48	83.33	8.34	1.85
[o]	0.00	1.85	0.93	70.85	20.37
[U]	0.00	0.00	0.00	5.55	94.45
				OVERALL	86.85

c) Area function - compressed and range normalized (AFCR)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	98.14	0.93	0.00	0.93	0.00
[e]	1.85	94.44	2.78	0.93	0.00
[ ^ ]	0.93	5.55	91.67	1.85	0.00
[o]	0.93	0.93	6.48	89.81	1.85
[U]	0.00	1.85	0.93	12.04	85.18
				OVERALL	91.85

**Table-5.3 : Performance of vowel discrimination - Speaker C**  
(Confusion matrix in percentage values)

**Female Voice**

a) Cepstral coefficients (CEP)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	98.33	1.67	0.00	0.00	0.00
[e]	2.00	98.00	0.00	0.00	0.00
[ ^ ]	0.00	0.00	100.00	0.00	0.00
[o]	0.00	0.00	8.00	84.00	8.00
[U]	2.15	0.70	1.67	16.66	81.67
OVERALL					92.40

b) Area function - compressed (AFC)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	98.00	2.00	0.00	0.00	0.00
[e]	4.00	96.00	0.00	0.00	0.00
[ ^ ]	0.00	0.00	97.92	2.08	0.00
[o]	0.00	0.00	0.00	70.00	30.00
[U]	0.00	0.00	0.00	22.00	78.00
OVERALL					87.98

c) Area function - compressed and range normalized (AFCR)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	98.00	2.00	0.00	0.00	0.00
[e]	8.00	92.00	0.00	0.00	0.00
[ ^ ]	0.00	0.00	100.00	0.00	0.00
[o]	0.00	0.00	0.00	80.00	20.00
[U]	0.00	0.00	4.00	18.00	78.00
OVERALL					89.60

**Table-5.4** : Performance of speaker independent vowel discrimination  
(Confusion matrix in percentage values)

**a) Cepstral coefficients (CEP)**

TEST	REFERENCE				
	[I]	[e]	[^]	[o]	[U]
[I]	96.43	0.00	0.00	0.00	3.57
[e]	27.86	70.00	0.00	0.00	2.14
[^]	0.00	1.43	98.57	0.00	0.00
[o]	0.00	0.00	8.57	89.29	2.14
[U]	0.00	1.43	0.00	15.71	82.86
				OVERALL	87.43

**b) Area function - compressed (AFC)**

TEST	REFERENCE				
	[I]	[e]	[^]	[o]	[U]
[I]	92.86	5.72	0.71	0.00	0.71
[e]	49.29	47.85	1.43	1.43	0.00
[^]	0.00	0.71	45.71	50.72	2.86
[o]	0.00	0.00	54.29	43.57	2.14
[U]	0.00	2.14	0.71	12.86	84.29
				OVERALL	62.86

**c) Area function - compressed and range normalized (AFCR)**

TEST	REFERENCE				
	[I]	[e]	[^]	[o]	[U]
[I]	100.00	0.00	0.00	0.00	0.00
[e]	67.86	30.71	1.43	0.00	0.00
[^]	0.00	1.43	97.14	1.43	0.00
[o]	0.00	1.43	27.14	62.86	8.57
[U]	0.71	3.57	5.00	15.00	75.72
				OVERALL	73.29

s study are shown in Table–5.5. The **results** were seen to be on par with the speaker dependent results indicating that the network can learn the features of both the speakers. It is seen that the range normalized area coefficients results in consistently poor performance for vowel [U] when compared to the compressed area coefficients.

Table–5.6 summarizes the performance results highlighting the comparison across different parametric representations. The cepstral representation does score well relative to the other representation. It is observed that the errors occur consistently with the neighbors for the area coefficients representation when testing across speakers. Table–5.7 summarizes the performance results highlighting the comparisons for different vowels. The **results** are more or less consistent for the different vowels except for the case when comparing across parameters.

The activation levels of the network outputs provide additional information. The extent of activation gives an indication of the confidence with which the classification occurs. The second best activation in the output suggests a possible **candidate** in case of errors. The **recognition** performance **considering** the **second** best activation in outputs is shown in Table–5.8. It is seen that there is significant improvement in the performance when the second best candidate is also included.

To summarize, the vowel classification **performance** in the vowel region of the CV utterance shows performance greater than 92% using the cepstral parameters. Between the compressed area coefficients and the range normalized area coefficients, the latter's performance is distinctly better except for one speaker's data. Even for this case, the degradation is mainly **due** to the poor score for the vowel [U]. The performance of area coefficients parameters improves to greater than **94%** when the second best candidate is also considered.

**Table-5.5** : Performance of multispeaker vowel discrimination  
(Confusion matrix in percentage values)

a) Cepstral coefficients (CEP)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	90.73	7.66	0.40	0.00	1.21
[e]	1.21	98.39	0.40	0.00	0.00
[ ^ ]	0.00	0.81	98.79	0.40	0.00
[o]	0.00	0.00	6.45	91.13	2.42
[U]	0.40	0.00	0.40	4.84	94.36
OVERALL					94.68

b) Area function - compressed (AFC)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	96.78	2.82	0.40	0.00	0.00
[e]	16.53	81.05	0.81	0.00	1.61
[ ^ ]	0.00	-2.82	81.05	13.31	2.82
[o]	0.00	0.81	6.05	73.79	19.35
[U]	0.00	0.00	0.40	6.05	93.55
OVERALL					85.24

c) Area function - compressed and range normalized (AFCR)

TEST	REFERENCE				
	[I]	[e]	[ ^ ]	[o]	[U]
[I]	91.13	7.66	0.40	0.00	0.81
[e]	8.06	90.73	0.00	0.00	1.21
[ ^ ]	0.00	3.23	89.92	5.24	1.61
[o]	0.40	2.42	6.45	77.82	12.90
[U]	0.00	1.61	0.81	8.47	89.11
OVERALL					87.77

**Table-5.6 :** Summary of correct classification performance for different parametric representations

	CEP	AFC	AFCR
<b>Speaker Dependent</b>			
• Speaker A	94.28	85.57	79.43
• Speaker B	95.93	86.85	91.85
• Speaker C	92.40	87.98	89.60
<b>Speaker Independent</b>	87.43	62.86	73.29
<b>Multispeaker</b>	94.68	85.32	87.82

**Table-5.7** : Summary of correct classification performance for different vowels

## a) Cepstral coefficients (CEP)

	[I]	[e]	[^]	[o]	[U]
<b>Speaker Dependent</b>					
* Speaker A	84.28	100.00	100.00	92.86	94.29
• Speaker B	98.14	95.37	97.22	92.37	93.52
* Speaker C	98.33	98.00	100.00	84.00	81.67
<b>Speaker Independent</b>	<b>96.43</b>	<b>70.00</b>	<b>98.57</b>	<b>89.29</b>	<b>82.86</b>
<b>Multispeaker</b>	<b>90.73</b>	<b>98.39</b>	<b>98.79</b>	<b>91.13</b>	<b>94.35</b>

## b) Area function - compressed (AFC)

	[I]	[e]	[^]	[o]	[U]
<b>Speaker Dependent</b>					
* Speaker A	84.28	99.30	70.72	87.14	86.43
* Speaker B	87.96	91.67	83.33	70.85	94.45
• Speaker C	98.00	96.00	97.92	70.00	78.00
<b>Speaker Independent</b>	<b>92.86</b>	<b>47.85</b>	<b>45.71</b>	<b>43.57</b>	<b>84.29</b>
<b>Multispeaker</b>	<b>96.77</b>	<b>81.05</b>	<b>81.05</b>	<b>73.79</b>	<b>93.95</b>

## c) Area function - compressed and range normalized (AFCR)

	[I]	[e]	[^]	[o]	[U]
<b>Speaker Dependent</b>					
• Speaker A	88.57	99.30	81.43	82.14	45.31
• Speaker B	98.14	94.44	91.67	89.81	85.18
• Speaker C	98.00	92.00	100.00	80.00	78.00
<b>Speaker Independent</b>	<b>100.00</b>	<b>30.71</b>	<b>97.14</b>	<b>62.86</b>	<b>75.72</b>
<b>Multispeaker</b>	<b>91.53</b>	<b>90.73</b>	<b>89.92</b>	<b>77.82</b>	<b>89.11</b>

**Table-5.8 : Vowel classification considering the best two candidate classes**  
**Speaker A**

**a) Cepstral coefficients (CEP)**

	[i]	[e]	[^]	[o]	[U]	Overall
<b>Best candidate</b>	<b>84.28</b>	<b>100.00</b>	<b>100.00</b>	<b>92.86</b>	<b>94.29</b>	<b>94.29</b>
<b>Best two candidates</b>	<b>94.29</b>	<b>100.00</b>	<b>100.00</b>	<b>96.43</b>	<b>100.00</b>	<b>98.14</b>

**b) Area function- compressed (AFC)**

	[i]	[e]	[^]	[o]	[U]	Overall
<b>Best candidate</b>	<b>84.28</b>	<b>99.30</b>	<b>70.72</b>	<b>87.14</b>	<b>86.43</b>	<b>85.57</b>
<b>Best two candidates</b>	<b>97.86</b>	<b>100.00</b>	<b>92.86</b>	<b>95.00</b>	<b>97.14</b>	<b>96.57</b>

**c) Area function- compressed and range normalized(AFCR)**

	[i]	[e]	[^]	[o]	[U]	Overall
<b>Best candidate</b>	<b>88.57</b>	<b>99.30</b>	<b>81.43</b>	<b>82.14</b>	<b>45.71</b>	<b>79.43</b>
<b>Best two candidates</b>	<b>96.43</b>	<b>100.00</b>	<b>92.86</b>	<b>90.00</b>	<b>92.14</b>	<b>94.28</b>



The main reason for focussing on area coefficients as parameters for the vowel discrimination is to have parameters close to articulatory description. This would help in the development of a speaker independent recognition, as the articulatory feature description is independent of speaker. But it turns out that extracting these and articulatory features from speech data is extremely difficult. The area coefficients derived from linear prediction analysis have some limitations due to large variability and dynamic range. Moreover, these coefficients are critically dependent on the accuracy of the linear prediction analysis. However, it is interesting to see that they do bring out the discriminability and also the confusion is among vowels having vocal tract shape close to each other.

Having seen the discrimination performance in the dominant vowel region (the region after vowel onset) of the CV utterance, we shall now see methods to carry out discrimination of the manner, using features from the region before vowel onset instant, in the next chapter.

## STUDIES ON RECOGNITION OF CONSONANT MANNERS

The consonant category of **CV** utterances of Hindi is predominantly determined by the speech signal characteristics in the region before vowel onset. The recognition of the broad class of the regions before vowel onset in a CV utterance can be viewed from different perspectives. One view point is that of determining one among different consonant types (manners), for which a pattern recognition approach **may be** used. The parameters of the relevant part of the utterance can be used **as** a signature for comparison. The problems that need to be addressed arise due to **variability** in speech and associated parameters.

Another viewpoint is that of determining the occurrence of regions using a feature detection approach. A combination of suitably weighted parameters can be used to detect regions with specific features in the signal. This approach calls for identifying for each type of region, the choice of parameters, weighting function, discrimination thresholds, and algorithm/logic for combining multiple parameters. Region characteristics may also include durational features for some cases. The parametric variations due to the variability in speech may call for the use of fuzzy discrimination thresholds. Alternatively, the fuzzy thresholds may be imbibed indirectly using approaches based on artificial neural networks.

Intuitively, the latter viewpoint seems to hold promise, because, there exist possibilities of using region specific features to dictate the signal processing as well as recognition methodology. The characteristics of different regions can be

obtained from acoustic-phonetic analysis. This basis allows one to derive and apply speaker-independent and speaker-invariant features.

In this chapter, we discuss various aspects of realizing discrimination of the different regions of consonants. We have seen in Chapter 3 that CV utterances of Hindi can be broadly classified into groups based on the consonant types. We have also seen that each of the consonant groups exhibit regions with distinct characteristics and that detection of the occurrence of regions can directly lead to recognition of the consonant group to which a particular utterance belongs. The speech production mechanism characteristics and the acoustic features in the different regions of the CV utterances are also highlighted in Chapter 3. With regions and their features as the starting point, the important features for discrimination of regions are evolved. This forms the acoustic-phonetic basis for discrimination. The important features for discrimination of the different regions leads to the discussion on the choice of parameters to detect features and then to the signal processing issues. This is followed by the discussion on the studies carried out to assess the behavior of parameters. Other more direct features pertinent to the discrimination of regions are also discussed. The discussions lead to a methodology for the discrimination of the consonant regions in CV utterances. The studies carried out and the performance realized using the proposed method are discussed in the concluding part of the chapter.

## **6.1 DISTINGUISHING FEATURES OF CONSONANTAL REGIONS**

The acoustic features and consequent waveform characteristics have been discussed in Chapter 3 for the three different region types in CV utterances. It is observed that features that may be termed as speaker specific characteristics are the actual values of some of the parameters. Some examples of speaker specific features are the pitch period and its variations, the specific glottal wave shape dictated by the speaker's glottal vibrations, and the frequency values of the gross

spectral peaks and **nulls** arising out of the specific dimensions of the vocal tract acoustic tube. Thus, to obtain speaker independent cues for discrimination, it is necessary to avoid the use of features which show significant **dependence** on speaker characteristics. Taking this aspect into account, the significant features for discrimination of the different region types, can be obtained by considering the following:

- perceptual features (phonetic aspects).
- waveform features (time domain),
- acoustic source features (nature of excitation), and
- acoustic system features (gross spectral features)

With these view points as basis, characteristic features under the above indicated broad categories of features which show distinct differences for different - regions are indicated in Table-6.1. The **perceptual** features (or rather the phonetic considerations) are not explicitly indicated. They are implicit in the broad classification of the utterances. The other features that have been considered show conspicuous presence or absence in different regions. **They** are mostly gross features implying averaged characteristics with minimal dependence on speaker traits. Waveform periodicity is one such feature as speech generally exhibits regions which have either periodic structure or random variations. Another significant feature is the relative energy levels in different regions of speech. This feature is mostly determined by the nature of the acoustic source, system and radiation characteristics. The different region types show distinct differences in this feature. **The** duration of different region types is another significant feature which is dependent on the nature of excitation and its sustainability from the point of view of production. Significant spectral structure is imposed on speech signals with excitation at an inner point in the vocal tract because of transmission properties of the acoustic tube. This structure is absent for speech signals with excitation close to the mouth. The absence of high frequencies in voicing, their presence in vowel-like regions and their predominance in fricative regions suggests

**Table-6.1:** Characteristics of features

**A: Waveform features:**

- a. Waveform periodicity (PW)
  - 1. Periodic
  - 2. Random
- b. Relative energy (EN)
  - 1. Low
  - 2. Significant
  - 3. High
- c. Duration (D)
  - 1. Short
  - 2. Long
  - 3. Very long

**B: Acoustic system features :**

- a. Spectral structure (SS)
  - 1. Absent
  - 2. Weak
  - 3. Strong
- b. High frequency energy (HF)
  - 1. Low
  - 2. Significant
  - 3. High
- c. Gross spectral nulls (SN)
  - 1. Absent
  - 2. Present

**C: Acoustic source features :**

- a. Periodicity of excitation (PE)
  - 1. Periodic
  - 2. Random
  - 3. Single impulse
- b. Strength of excitation (SE)
  - 1. Weak
  - 2. Significant
  - 3. Strong
- c. Average count of excitation (NE)
  - 1. One
  - 2. Low
  - 3. High

the presence or absence of high frequencies as an important feature. Frequency selective absorption due to coupled acoustic cavities is present in semivowels and nasals and this leads to gross spectral nulls in these regions. The difference in region types is a direct **consequence** of the nature of acoustic **excitation/source**. In speech, the acoustic excitation can be broadly identified as being either periodic impulses, or random noise-like excitation or single impulse (shock) excitation. The effective strength of this excitation is dependent on its nature as well as the acoustic system (vocal tract transmission and radiation) characteristics.

Another feature that is useful but not included in the Table-6.1 is the energy variance in the region. Pitch excited regions show very low energy variance, whereas, release transient, aspiration and voiced aspiration regions exhibit significant energy fluctuations.

Table-6.2 lists the characteristics of features in different regions. It is observed from this table that, some of the features seem to be redundant as they exhibit identical trends for different regions. For example, the features - spectral structure (SS), periodicity of excitation (**PE**) and count of excitation in region (NE) - discriminate the same group of regions. This does not mean that these features are identical. What is implied is that these different features are manifestations of a common production mechanism characteristics. The redundancy can be viewed as providing reinforcement to the discrimination process. This is necessary in the context of speech as it is difficult to compartmentalize any feature into the indicated categories because of the variability in speech. Table-6.3 combines these features and lists the distinguishing features between region pairs. Entries in the table indicate that the corresponding feature is different for the pertinent region pair. Absence of entries for semivowel - nasal region pair implies that there are no gross features that can discriminate between them. In other words, discrimination between these region types calls for more detailed features to be considered. All

**Table-6.2 : Feature characterization in different regions**

	FEATURE CHARACTERISTICS																								
	PW		EN			D			SS			HF			SN		SE			PE			NE		
REGION	1	2	1	2	3	1	2	3	1	2	3	1	2	3	1	2	1	2	3	1	2	3	1	2	3
VNG		*		*			*		*			*			*			*		*			*		
RLT	*			*		*			*			*			*			*		*			*		
ASP	*				*		*		*			*			*			*		*			*		
VAS		*			*		*		*			*			*			*		*			*		
NSL		*		*		*			*			*			*			*		*			*		
SMV		*		*		*			*			*			*			*		*			*		
FRC	●		●			*			*			●	●		●			●		●			●		
SIL	*		●				●	●				●	*		●			●		●			●		

VNG : Voicing

RLT : Stop release transient

ASP : Aspiration

VAS : Voicedaspiration

NSL : Nasal

SMV : Semivowel

FRC : Fricative

VWL : Vowel

Note : Characterization of features are as indicated in Table-6.1.

**Table -6.3 : Distinguishing features of regions**

	VNG	RLT	ASP	VAS	NSL	SMV	FRC
SIL	$\begin{bmatrix} * & & * \\ & * & \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} & * & * \\ & * & \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} & * & * \\ * & * & \\ & * & \end{bmatrix}$	$\begin{bmatrix} * & * & * \\ * & * & \\ & * & \end{bmatrix}$	$\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} * \\ \\ * \end{bmatrix}$
FRC	$\begin{bmatrix} * \\ * & * & * \\ * & * \end{bmatrix}$	$\begin{bmatrix} * & * \\ * & * & * \\ * & * \end{bmatrix}$	$\begin{bmatrix} * \\ * & * \\ * \end{bmatrix}$	$\begin{bmatrix} * & * \\ * & * \\ * \end{bmatrix}$	$\begin{bmatrix} * & * \\ * & * & * \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} * & * \\ * & * & * \\ * & * & * \end{bmatrix}$	
SMV	$\begin{bmatrix} * \\ * & * & * \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} * & * \\ * & * \\ * & * \end{bmatrix}$	$\begin{bmatrix} * & * \\ * & * \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} * \\ * & * \\ * & * & * \end{bmatrix}$			
NSL	$\begin{bmatrix} * \\ * &   & * \\ * &   & * \end{bmatrix}$	$\begin{bmatrix} \bullet & * \\ \bullet & * \\ \bullet & * \end{bmatrix}$	$\begin{bmatrix} * & * \\ * & * \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} * \\ * & * \\ * & * & * \end{bmatrix}$			
"AS	$\begin{bmatrix} * \\ * & * \\ * & * \end{bmatrix}$	$\begin{bmatrix} * & * & * \\ * \\ * & * & * \end{bmatrix}$	$\begin{bmatrix} * \\ \\ \end{bmatrix}$				
ASP	$\begin{bmatrix} * & * \\ * & * \\ * & * \end{bmatrix}$	$\begin{bmatrix} * & * \\ * \\ * & * & * \end{bmatrix}$					
RLT	$\begin{bmatrix} * & * & * \\ * \\ * \end{bmatrix}$						

**Note 1 :** The above matrices refer to the feature matrix given below

PW	EN	D
SS	HF	SN
PE	SE	NE

**Note 2 :** Presence of asterisk (\*) Indicates that the feature specified in that position is a distinguishing feature for the region-pair.



other regions have **atleast** one or more features which differ between each region pair.

Having identified features for discrimination of regions, it is necessary to obtain parameters by signal processing methods which can bring out the relevant gross features from the speech signal. This is the topic for discussion in the next section.

## **6.2 CHOICE OF PARAMETERS FOR CONSONANT REGION DISCRIMINATION**

Parameters provide the link between the speech signal and the observed features. Features as we have seen are usually qualitative. Features represent manually and visually observed characteristics or patterns. They embody complicated decisions on the importance of a feature in context. Parameters on the other hand provide quantitative measurements of some signal characteristics but lack the mechanism for discerning the patterns. The aim is thus to obtain suitable parameters which reflect the feature traits, or, viewed differently, to determine the efficacy of a parameter in highlighting a feature. We shall now consider parameters which can bring out each feature. Unless otherwise indicated, the parameters are normally obtained through short-time (10msec) analysis of speech data.

### **6.2.1 Waveform periodicity (PW)**

In detecting waveform periodicity, we are interested not in the period but in the feature that the waveform exhibits pattern similarity. A suitable parameter which can detect the occurrence of this similarity is a correlation parameter such as autocorrelation considering different lags. The correlation would maximize when the lag equals the period. A simpler version of this correlation can be obtained by considering average magnitude difference function (AMDF) [Ross 1974]. The AMDF exhibits a significant minimum when the lag approaches the period. Since

the AMDF would exhibit arbitrary minimum for any signal, it is necessary to use a suitable threshold, both in magnitude as well as in the range of tags to detect that the minimum is that due to expected periodicity. The magnitude threshold can be made dependent on the signal level by using the average magnitude of the signal [Jayant 1984]. Detection of periodicity using this approach is hindered by the period to period variations that occur in natural speech signal. Another problem is that due to the nature of the region. Typically, the consonantal regions are relatively low in energy. Consequently, background noise can play a significant role in dictating the performance. In such a situation, rather than AMDF, the lagged autocorrelation peak serves as a more robust parameter. Measurement of average zero-crossings count of the signal can also yield some measure of the periodicity in the signal. The count is low for periodic signals, whereas for random noise like signals, the count is high. To avoid problems due to background noise corrupting the measurement, in practice, suitable threshold crossings are counted rather than zero-crossings. More elaborate period detection and measurement algorithms can also be used [Markel 1973]. But the computational burden is generally high in these methods.

### **6.2.2 Relative energy (EN)**

The short-time signal energy can be obtained by straight forward computations from the signal. The relative energy across regions can be obtained by considering the normalized energy (relative to the maximum in the region of interest in the signal).

### **6.2.3 Duration of region (D)**

The duration feature is obtained as a direct parameter when considering regions with manually identified boundaries during analysis. To obtain this as a parameter in a recognition / discrimination context calls for the identification of region boundaries with reasonable resolution and correctness.

#### **6.2.4 Spectral structure (SS)**

The spectral structure as a feature refers to the occurrence of gross spectral emphasis due to formants. A parameter that can bring out this feature is the spectral flatness measure obtained as a byproduct of linear predictive modeling of the speech signal [Gray 1974]. The measure gives an indication of how well the model spectrum fits the signal spectrum. Regions with significant spectral structure are modeled well resulting in low value of spectral flatness, whereas, regions which have noise like behavior are not modeled well and exhibit spectral flatness measure close to unity. Regions in the signal exhibiting single gross peak and large spectral dynamic range also result in being modeled well though in the sense of spectral structure as construed by us it does not exhibit any spectral structure. It may be more appropriate in such cases to pre-flatten the signal spectrum using suitable first order correction before determining spectral flatness.

#### **6.2.5 Relative high frequency energy (HE)**

This feature may be realized through several parameters. The first linear prediction coefficient is a good approximation. The variability in speech poses a problem in defining the threshold frequency separating low and high frequencies. Barring this, it is possible to obtain relative high frequency energy as a parameter by directly taking the spectral energy of the signal in the low and high frequency regions. The term 'relative' in our usage refers to the feature's behavior across the utterance. Another parameter which can provide an approximation of this feature is the average count of zero crossings in the signal. This parameter has the advantage that it is computationally simpler.

#### **6.2.6 Gross spectral nulls (SN)**

This feature calls for determining the occurrence of regions in the signal exhibiting gross dips (zeros) in the spectrum due to spectral absorption. Linear predictive modeling is not useful in this context as it normally fits only the spectral peaks. But, it is possible to apply it on a suitably inverted spectrum. Another

approach is to use sophisticated pole-zero decomposition methods [Yegnanarayana 1981].

### **6.2.7 Summary**

Since several parameters are identified which provide a certain approximation of the feature, a suitable set is considered for study. The parameters used in the study are listed in Table-6.4. The trend parameters are based on the change in the parameters from the initial part to the final part of the region. The histograms of some of these parameters in different type of regions for one set of utterances are shown in Fig. 6.1. The histograms show that, individually, each parameter exhibits distinct overlap in parametric value for the different regions. Yet, it is possible to realize discrimination by combining evidences from several parameters.

Since it is known that the consonantal region typically consists of weak signal, spectral parameters, even if they are important for some cases, cannot be relied upon due to sensitivity of the spectral parameters to the inevitable additional noise in these regions. Hence use of detailed spectral parameters were consciously avoided.

## **6.3 CLASSIFICATION OF CONSONANTAL REGIONS OF CV UTTERANCES**

To realize discrimination of the consonant manner, we have taken recourse to the use of artificial neural network classifier. The use of neural network approach makes it possible to combine the different parameters in a convenient manner. The parameters are suitably normalized before being fed as input to the neural network. Even if some parameters do not bring out the discrimination well, the training procedure will result in assigning a low weightage to the corresponding input. The parameters were derived from manually identified regions in the training data set. Likewise, even the test data was used with manually identified region

**Table-6.4 :** Parameters for region classification

1. Normalized energy
2. Normalized average magnitude
3. Average zero-crossings count
4. Normalized energy of differenced signal
5. Normalized average magnitude of differenced signal
6. Average zero-crossings count of differenced signal
7. Duration
8. Spectral flatness measure
9. Lagged autocorrelation peak
10. Lag of autocorrelation peak
11. First order LP coefficient using autocorrelation value from lagged peak
12. Trend in parameter 10
13. Variance of LP residual
14. First order LP coefficient
15. Trend in parameter 1
16. Trend in parameter 2
17. Trend in parameter 3
18. Trend in parameter 4
19. Spectral distance between initial and final parts of the region (Itakura measure)

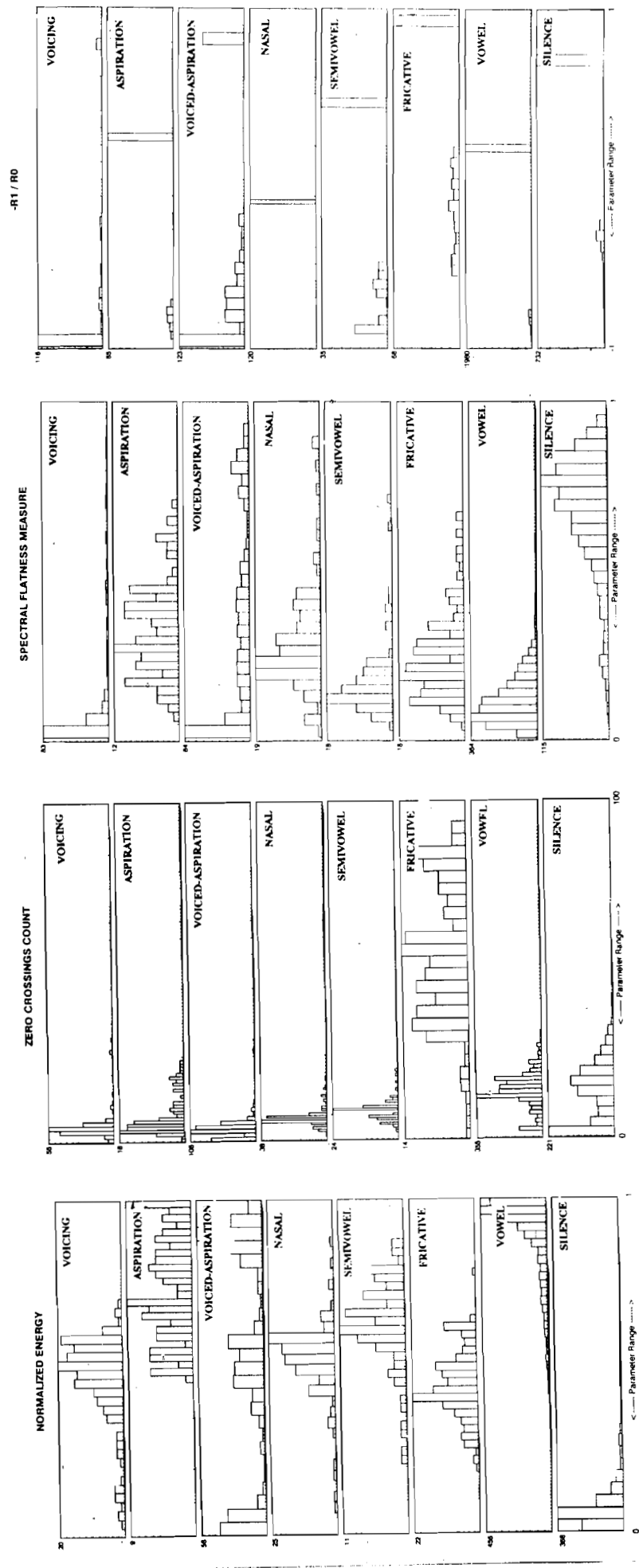


Fig. 6.1 : Histograms of some parameters in different regions of CV utterances.

boundaries. As discussed, the region identified for analysis is that prior to vowel onset in the CV utterance (the region between start of utterance and vowel onset). Since this region spans widely varying duration for the different manners, the average and trend of parametric behavior in the region is also determined and used. The duration of the region is an **important** input and this is determined knowing the instants of utterance start and vowel onset.

A feed forward network with two hidden layers was used. The number of inputs is dependent on the number of parameters chosen for representation. The studies were carried out with different sets of parameter. The choice of parametric representation is crucial since the manner classification needs to be performed in any vowel context. The number of nodes in the hidden layers was chosen taking into consideration the number of input nodes. (In cases when the number of parameters in the experiment is large, the number of nodes in the hidden layers can be less whereas, in cases when input parameters are less the number of nodes in the hidden layers is made large.) The number of output nodes of the classifier is directly dependent on the number of manner classes with one output node identified for each class. The nasal consonants were not included in the study and ,therefore the number of manner classes is six. The actual network size used in each study is indicated in the tables.

## **6.4 DISCUSSION OF RESULTS**

The studies were carried out with data obtained for a single speaker. Table-6.5 shows the confusion matrix realized for one speaker's test data with the first fourteen gross parameters listed in Table-6.4. Table-6.5 is arranged such that the neighboring consonant manners in the table are closer in terms of features. Region pairs which are closer in terms of features can be identified from Table-6.3 as those with less number of entries in the matrix. If the entries are few in number, it implies that the distinguishing features between them are few.

The results in Table-6.5 show an overall classification performance of only 60%. Only one manner shows high score namely that of voiced-aspirated. The fricatives and semivowels have shown distinctly poor performance. To see if the cause for this poor performance was the variability in some of the parameters, a reduced set consisting of only the first ten parameters which are relatively robust were used. The results of this study are shown in Table-6.6. A marginal improvement in overall performance is seen (63%). A detailed inspection of Tables-6.5 and 6.6 show that there is distinct improvement in the classification of three of the manners - semivowel (45% to 59%), voiced-aspiration (85.65% to 92%) and aspiration (64.8% to 85.6%), but at the cost of increased confusion in voiced-unaspirated and unvoiced- unaspirated manners. The fricative manner shows marginal improvement. Since in both the studies only averaged parameters of the region were used, it was felt that inclusion of parametric trend could reduce some of the confusions. Trends in the parameters, as indicated in Table-6.4 entries 15 to 18 were also included. A spectral distance parameter was also included to qualify the gross trend in spectral features in the -consonantal region. In all, nineteen parameters (listed in Table-6.4) were used in the third study. The performance using these nineteen parameters is shown in Table-6.7. A distinct improvement in performance was observed in the overall performance (69%). There was marked improvement in the classification of voiced- unaspirated (34% to 58%), semivowel (59% to 77%) and unvoiced- aspirated (86% to 90%) manners. The voiced-aspiration remained at the same level of 92%. A marginal improvement in the fricative classification was also observed (57% to 61%). The only degradation is for the unvoiced-unaspirated (RLT) which shows decrease in performance (48% to 35.2%). The poor performance can be attributed to the fact that the unvoiced-unaspirated manners have a very short duration of the consonantal region. Consequently, the gross parameters which were based on block data processing do not reflect the nature of the region well. As pointed out earlier, neighboring columns in the table refer to manners which are closer in terms of features. It is seen from Table-6.7 that the confusions are typically with the



**Table-6.5** : Manner classification performance

TEST	REFERENCE					
	VNG	SMV	VAS	ASP	RLT	FRC
VNG	53.60	12.00	11.20	08.80	07.20	07.20
SMV	11.00	45.00	28.00	12.00	01.00	03.00
VAS	00.00	11.20	85.60	03.20	00.00	00.00
ASP	04.80	08.00	08.00	64.80	01.60	12.80
RLT	11.20	14.40	03.20	06.40	56.00	08.80
FRC	-14.00	04.00	00.00	14.00	14.00	54.00
					OVERALL	60.57

**Table-6.6** : Performance of manner-classification using reduced set of parameters

TEST	REFERENCE					
	VNG	SMV	VAS	ASP	RLT	FRC
VNG	33.60	08.80	04.80	28.00	17.60	07.20
SMV	10.00	59.00	10.00	18.00	02.00	01.00
VAS	00.80	01.60	92.00	02.40	01.60	01.60
ASP	00.80	00.00	02.40	85.60	09.60	01.60
RLT	13.60	00.80	00.80	26.40	48.00	10.40
FRC	11.00	02.00	01.00	18.00	11.00	57.00
					OVERALL	62.86

adjoining columns in the table. This can be clearly seen for the semivowels, voiced-aspiration, aspiration and fricatives. To some extent, the unvoiced-unaspirated and voiced-unaspirated region also show maximal confusion with the neighbours. This suggests the possibility that the parameters chosen do not adequately reflect the required distinction between the close members. This implies that more sophisticated parameters need to be evolved which represent the distinguishing features better.

It was observed that the consonantal part shows features dependent on the vowel context. The variability in the parameters due to vowel context was considered as another possibility for the confusion. A fourth study was therefore conducted to see if the performance improved when the consonants with the same vowel context were grouped and tested using all nineteen parameters. The results of this study are shown in Table-6.8. Each column in the table refers to the indicated vowel context of the CV utterance used in the study. The rows of the table show the classification performance for each of the manners. The combined overall performance does not show any improvement compared to Table-6.7. This suggests that the confusion in classification is more due to the inadequacy of the parametric representation rather than the variability due to the vowel context. But more detailed comparison within the table shows that some of the consonant manners are better resolved when compared with the earlier study. The results in Table-6.8 indicate a definite improvement in performance for the unvoiced-unaspirated regions in all the five vowel contexts.

In summary, the overall manner classification realized is 69%. This is obtained after including trend and distance parameters. Inclusion of vowel context has not shown much improvement in performance. The trend in performance in different stages of study suggests that further improvement can be obtained by refining the parameters. Having looked at approaches for determining the

**Table-6.7 :** Performance of manner classification including trend parameters  
(Confusion matrix in percentage values)

TEST	REFERENCE					
	VNG	SMV	VAS	ASP	RLT	FRC
VNG	58.4	9.6	0.8	13.6	6.4	11.2
SMV	30.0	77.0	9.0	7.0	2.0	2.0
VAS	0.8	4.0	92.0	2.4	0.0	0.8
ASP	1.6	0.0	3.2	90.4	0.8	4.0
RLT	12.8	1.6	0.8	28.0	35.2	21.6
FRC	0.0	3.0	0.0	32.0	4.0	61.0
					OVERALL	69.1

**Table-6.8 :** Performance of manner classification in the context of different vowels  
(Confusion matrix in percentage values) -

MANNER	VOWEL CONTEXT				
	[i]	[e]	[^]	[o]	[U]
VNG	56.00	64.00	64.00	56.00	64.00
SMV	80.00	60.00	60.00	55.00	45.00
VAS	92.00	64.00	52.00	96.00	80.00
ASP	68.00	60.00	76.00	86.00	48.00
RLT	52.00	60.00	60.00	64.00	64.00
FRC	60.00	70.00	85.00	80.00	45.00
OVERALL	67.00	62.90	65.70	72.80	58.60
				OVERALL	65.60

---

consonant manner, we shall address the issue of determining place of articulation in the next chapter.

## **STUDIES ON RECOGNITION OF PLACE OF ARTICULATION**

As discussed in Chapter 3, the place of articulation refers to the place in the vocal tract where release of constriction occurs in the production of the consonant. The manner of the consonant defines the context of this release. In nasals, semivowels (with the exception of alveolar-trill) and fricatives the articulatory **release** does not induce any new source of signal. It merely causes a change in the shape of the vocal tract. Specifically for the fricatives, a change in the type of excitation also occurs around the same instant. On the other hand, since stop release results in abrupt release of air pressure at the place of **stricture**, there is an impulse excitation of the vocal tract at the instant of release as well as random excitation due to turbulent air flow for a short time when the vocal tract is constricted during the transition from closed to open. The nature of the resultant signal for the stop releases (usually called the burst or release transient) is characterized by the place of articulation and the shape of the vocal tract outwards from the place. But, the signal is present for a short ( $< 10$  msec) duration in most cases and is relatively weak as well. Hence, it is not always possible to obtain the needed cues from this short duration weak signal. This chapter discusses issues in analyzing the dynamic regions of CV utterance. It also considers the need for choosing appropriate parameters for determining the place of articulation in isolated CV utterances. A methodology for determining the place feature is suggested and the results obtained by using this methodology are also explained.

## 7.1 ISSUES IN ANALYZING THE DYNAMIC REGION

Since the movement of articulators spans a finite time due to the dynamics of the production mechanism, the vocal tract exhibits shape change following the release. We call this dynamic region as the transition region. Except for the case of aspirated stops, the **source** of excitation in this transition region is invariably the strong glottal vibrations which results in high SNR signal. This high SNR region is amenable for spectral analysis. Therefore, a more reliable cue to the place of articulation is obtained by analyzing the signal in this transition region for characterizing the vocal tract shape change. It is to be noted that the time span of the articulatory movement during this consonantal release is **mainly** due to the natural dynamics of the articulators and is to a significant extent not under the control of the speaker. This is in contrast to the signal produced under steadily maintained positions of the articulators which can be prolonged or shortened under speaker's control. Thus, the durational variability is significantly less in this transition region. This makes it possible to analyze a fixed duration of the signal after the instant of release to characterize the transition.

Even in the case of consonants other than stops, the analysis of the transition region provides good cues to determining the place of articulation. However, for the cases of nasals and semivowels, relatively strong signal is present prior to consonant release. This is because, glottal vibrations are initiated in advance and free air flow is possible through the vocal tract due to partial opening. Thus, the signal in the region before release is characterized by the shape of the vocal tract dictated by the consonantal stricture. It is possible to make use of this cue as well for these consonants.

Analogously, the fricative consonants exhibit random-noise like excitation in the region before release. The gross spectral characteristics in this noise-like region is dictated by the place of stricture or rather the vocal tract outwards from

the place of stricture. But, the signal is relatively weak in this region. Hence one can rely on this feature as a cue to the place of articulation only to a limited extent.

For the aspirated stops, the transition region occurs completely in the initial part of the aspiration region. The aspiration region is characterized by noise-like excitation due to turbulent air flow. Consequently the signal is also relatively weak. But the signal has significant spectral structure as the random noise-like excitation is subject to filtering by the vocal tract. It is from this signal that transition characteristics need to be determined.

Thus, the main issue in determining the place of articulation is processing of the speech signal in the transition region to obtain the specific place in the vocal tract where the shape changes from closure to openness occurs. The processing is complicated by the short (25 msec) duration of the transition region as well as the differing contexts of the consonant class and the vowel class in which it can occur. For each consonant manner there are either four or five possible places of articulatory release and the release can occur in the context of any of the ten vowels of Hindi. The problem is compounded because, for the same consonant manner, some of the places of articulation are at close positions in the vocal tract. The vowel context plays a dominant role since the vocal tract shape as well as its change is determined by it. As discussed before, for some consonant manners, it is possible to obtain features of the place of articulation by analyzing the signal in the region before the release, though the signal is generally weak in this region.

## **7.2 CHOICE OF PARAMETERS TO DETECT PLACE FEATURES**

It has been customary [Harrington 1988] to characterize the place of articulation of stop consonants by the spectral characteristics of the release

transient (burst). But the burst is a short ( $<10$  msec) duration weak signal and, in addition, exhibits significant variability due to the consonant class and the vowel context. Hence, as further evidence of the place of articulation conventionally the spectral changes of the dynamic transition region have also been used. This change is characterized by the nature of movement of gross spectral peaks or rather the formants [Kewley-Port 1982], [Sundar 1987], [Raman 1985] and [Lieberman 1956]. Since formants are directly attributable to the shape of the vocal tract acoustic tube, they do provide the necessary cue. But, the values of formants, as discussed in the studies on vowel classification, are dependent on the specific speaker's vocal tract dimensions. Hence variability due to speaker characteristics is implicit when using formant transitions as cues for the place of articulation. The formant transitions are a function of the consonant class, the vowel context and the place of articulation. The above viewpoint holds good for consonants other than stops as well, except that, in lieu of the release transient, the signal region prior to release is characterized.

Since the determination of place of articulation is a crucial component in our studies, the use of transition cues become **important** to realize reasonable performance. Instead of determining transitions in formants, we have attempted the use parameters related to the articulatory shape transition directly. The area coefficients determined from the signal were used as parameter to qualify the vocal tract shape. It is necessary to obtain the change in the vocal tract shape to determine the place of articulation. To obtain this, the area coefficients were determined at short (5 msec) intervals in a fixed duration of the utterance immediately following the articulatory release instant. Since the relevant region is that immediately following vowel onset instant in all CV utterances except for the aspirated and voiced aspirated manners, the signal strength is high. Therefore, it is possible to obtain the spectral features reliably. Along the lines of study on vowel classification, weighted cepstral parameters were also used for comparison.



### 7.3 METHODOLOGY FOR RECOGNITION

Since the transition cues are defined by the specific place of articulation and the following vowel context, it is necessary to group the utterances so that discrimination is carried out amongst members of a subset of CV utterances. The specific places of articulation and the corresponding transition cues are dependent on the manner of consonant as well. It is assumed that the vowel context of the CV utterance as well as the consonant manner are known apriori. It is also assumed that the articulatory release instant is known. The articulatory release in the case of aspirated and voiced–aspirated manners occur along with the start of aspiration. Therefore the region for analysis is predetermined. The parameters extracted from the speech signal in this region are used in the context of known consonant manner and vowel class to determine the place of articulation. It is possible to combine some of the manner classes into a single group if the manner features do not affect the transition region features significantly. This is true of the manners like unvoiced–unaspirated stops, voiced–unaspirated stops, semivowels and fricatives. But the places of articulation are distinctly different for the fricative and semivowel class as compared to the stop consonants (See Table–3.1). Therefore in our studies, the unvoiced–unaspirated stops and the voiced–unaspirated stops alone are clubbed together. For the other manner classes – nasals, aspirated stops and voiced–aspirated stops, the transition region features are significantly affected by the manner features. This can be appreciated by seeing that the nasals exhibit nasalization of the transition region, while the aspirated stops as well as voiced–aspirated stops exhibit aspiration and voicing mixed with aspiration respectively, in the transition region.

Neural network classifiers with structure similar to that used for vowel classification, were used for this study also, for identical reasons. The main distinction is that the network has to handle more number of inputs for place discrimination because parameters from successive frames have to be used

simultaneously. The number of inputs is determined by the number of frames for representing the transition and the number of coefficients in each frame. Since spectral analysis here is identical to that performed for the vowel classification, the number of coefficients is 10 per frame. The number of frames of transition is determined by the duration of the transition region. Spectrographic studies indicated that the transition region usually spans a duration of 25 ms. The time resolution of the parametric representation was chosen to be 5 ms. This suggests that about 5 frames should suffice to characterize the changes in the transition region. Therefore, the number of inputs to the classifier was typically 50. The first hidden layer was chosen to have 40 nodes and the second hidden layer 20 nodes. The number of outputs corresponds to the number of places of articulation in each group. This is four for the semivowel and fricative groups and five for the others.

#### **7.4 STUDIES ON THE DETERMINATION OF PLACE OF ARTICULATION**

Different networks were used for each combination of vowel class and consonant manner. As discussed in the preceding section, appropriate manner classes are included in the same set to reduce the number of networks. Thus, there are five different manner classes, namely, aspirated, voiced aspirated, other stops, semivowels and fricatives. The number of vowel classes are five. Therefore, twenty five different networks were used corresponding to each combination of manner class and vowel class.

As was done for vowel classification the same three parametric representations were used in this study also. A preliminary study was carried out with number of frames ranging from 4 to 8. The results of classification indicated no marked changes from 4 to 6 frames but for the context of 7 and 8 frames there was distinct degradation in performance. The degradation is attributable to the undue emphasis of steady vowel features from the later frames. (The steady

vowel feature is nearly the same for different places of articulation since the vowel context is the same.) Therefore, in subsequent studies, the number of frames was chosen to be five.

## **7.5 DISCUSSION OF RESULTS**

The recognition performance for the different places of articulation is shown in Table-7.1. Since there are seventy five networks, considering all the three parametric representations, the consolidated performance is indicated in Table-7.1. Our conclusions based on the detailed results are given below as an embellishment to the results shown in Table-7.1.

It is seen that there is reduction in performance for area function as compared to the cepstral parameters, but the performance for the different places of articulation follow nearly the same trend. It is noticed that the different manners of stop consonants show nearly the same overall performance. The semivowels and fricatives exhibit high recognition performance with the only exception being palatal fricative. The confusion in this was noticed to be mostly with the adjacent place, retroflex and rarely with the next nearer place alveolar. It is observed that for the stop consonants also, most often, the confusion is with the neighboring places of articulation. Between the compressed area coefficients and the range normalized area coefficients, the latter shows better performance.

In summary, an overall performance greater than 90% is obtained for the classification of place of articulation. Considering that the place of articulation is exhibited as a weak feature, the high performance indicated shows that the approach used does provide the cues for place of articulation. Confining our attention to the region in the utterance exhibiting the articulatory shape change characterizes the place of articulation well enough to achieve good performance. The area functions have been used with a view to capture the articulatory change

Table-7.1 : Performance of classification of place of articulation.

Correct classification performance for different places of articulation in the context of different consonant manner in percentage values.

MANNER	PARAMETRIC REPRESENTATION	PLACE OF ARTICULATION					OVERALL
		VELAR	PALATAL	RETRO FLEX	DENTI-ALVEOLAR	BILABIAL	
UV-UA & V-UA	CEP	82	98	86	88	92	89.2
	AFC	70	88	68	82	78	77.2
	AFCR	76	86	72	82	80	79.2
ASP	CEP	96	91	83	87	81	87.6
	AFC	84	76	80	76	72	77.6
	AFCR	88	81	82	74	70	79.0
VAS	CEP	90	91	86	85	91	88.5
	AFC	84	88	72	76	76	79.2
	AFCR	83	90	78	75	79	81.0

MANNER	PARAMETRIC REPRESENTATION	PLACE OF ARTICULATION				OVERALL
		PALATAL	ALVEOLAR (TRILL)	ALVEOLAR (LATERAL)	LABIO DENTAL	
SMV	CEP	92	96	88	99	93.7
	AFC	92	74	84	96	86.5
	AFCR	90	78	82	98	87.0

MANNER	PARAMETRIC REPRESENTATION	PLACE OF ARTICULATION				OVERALL
		ALVEOLAR	RETROFLEX	PALATAL	GLOTTAL	
FRC	CEP	76	92	98	99	91.2
	AFC	72	84	88	96	85.0
	AFCR	70	86	87	97	85.0

directly. The performance though lower compared to cepstral coefficients, is high enough to justify its use. Keeping in mind that the area function exhibited large variability, the good recognition performance suggests that in spite of the variability in speech the dynamic characteristics are captured well in the gross shape of the function.

Having looked at the classification of the place of articulation in the context of the vowel and consonant manner, and at the independent classification of vowel and consonant manner in Chapters 5 and 6 respectively, it is necessary to see how the overall CV utterance can be determined. A scheme for recognition of overall CV utterances from the outputs of the classifiers for the vowel, consonant manner and place of articulation is discussed in the next chapter.

## STUDIES ON OVERALL RECOGNITION OF CV UTTERANCES

In this chapter we discuss studies to obtain the performance of a system for recognition of isolated utterances of **CV** units. We propose the need for a hierarchical model for classification of **CV** units and estimate the recognition accuracy that can be obtained from such a system. The estimated overall recognition is over 65%. Generally the confusing cases can be resolved using contextual information if the utterances are in the form of words or sentences although in each utterance the character is uttered in isolation. This is because most of the time the confused class is close to the true one from speech production point of view as seen from the results in the previous chapters for the first and the first two choices.

The starting point for the study is the collection of several samples of the isolated utterances of **CV** characters of all classes (total 140). Five samples were collected for each **CV** class, three were used for training and two for testing. Each utterance was segmented into three regions corresponding to manner, place and vowel. Parameter or features relevant to each region were extracted as discussed in the previous chapters. Recognition studies are conducted using each region in isolation and along with other regions for identification of class of each region.

It is not possible to have a single network for all the classes as the number of **CV** classes are large (140) and the number of training samples may not be adequate to train such a network. Moreover there will be confusion among several

classes due to closeness in the speech production of consonant manners, places of articulation and vowel types, especially for the places of articulation. Section 8.1 gives a summary of studies for the classification of each region data carried out separately in chapters 5, 6 and 7. In section 8.2 we consider the recognition of the class of a region using the parameters from the corresponding region as well as from an adjacent region. In section 8.3 we discuss studies on the recognition of the class of each region using parameters from all the three regions in the CV utterance. Section 8.4 gives a comparison of recognition results for each type, namely, manner, place and vowel. A hierarchical model of recognition system is proposed in section 8.4 and the performance estimated from such a model is discussed. Finally in section 8.5 we discuss the utility of these results and issues to be addressed for improving the recognition accuracy.

The network structure used in each of the studies is indicated in the corresponding table. The four values shown for the network structure correspond to the number of inputs, number of nodes in the first hidden layer, the number of nodes in the second hidden layer and the number of output nodes in that order.

## **8.1 CLASSIFICATION STUDIES OF EACH REGION USING PARAMETERS DERIVED FROM THE REGION ALONE**

In chapters 5, 6, and 7 we have studied the classification performance for the three regions in the CV utterance. Each of these studies had used parameters obtained from the corresponding region only. Classification results for these three regions are collected together and given in Table 8.1. An important assumption was that the region for analysis is known apriori. The neighboring regions' classes (manner and vowel) were assumed to be known for studies on classification of place of articulation.

**Table-8.1 :** Summary of the classification performance for the three parts of the CV utterance

a) Manner classification using nineteen gross parameters

The network structure is (19, 40, 25, 6)

TEST	REFERENCE					
	VNG	SMV	VAS	ASP	RLT	FRC
VNG	58.4	9.6	0.8	13.6	6.4	11.2
SMV	30.0	77.0	9.0	7.0	2.0	2.0
VAS	0.8	4.0	92.0	2.4	0.0	0.8
ASP	1.6	0.0	3.2	90.4	0.8	4.0
RLT	12.8	1.6	0.8	28.0	35.2	21.6
FRC	0.0	3.0	0.0	32.0	4.0	61.0
OVERALL						69.1

b) Classification of place of articulation using five frames of cepstral coefficients each from the transition regions

The network structure is (50, 40, 20, 5/4)

MANNER	PLACE OF ARTICULATION					OVERALL
	VELAR	PALATAL	RETROFLEX	DENTI-ALVEOLAR	BILABIAL	
UV-UA & V-UA	82	98	86	88	92	69.2
ASP	96	91	83	87	81	87.6
VAS	90	91	86	85	91	88.6
	PALATAL	ALVEOLAR (TRILL)	ALVEOLAR (LATERAL)	LABIO DENTAL		
SMV	92	96	88	99	93.7	
	ALVEOLAR	RETROFLEX	PALATAL	GLOTAL		
FRC	76	92	98	99	91.2	

c) Vowel recognition performance using ten cepstral coefficients from vowel region

The network structure is (10, 30, 20, 5)

TEST	REFERENCE				
	[i]	[e]	^	[o]	[U]
[i]	84.3	12.9	0.0	0.0	2.9
[e]	0.0	100.0	0.0	0.0	0.0
^	0.0	0.0	100.0	0.0	0.0
[o]	0.0	3.6	2.1	92.9	1.4
[U]	0.0	0.7	0.0	5.0	94.3
OVERALL					94.3



The above studies of the place of articulation used the context of vowel and manner of the utterance. That is place classification was performed independently for each combination of vowel and manner class. It is useful to obtain an idea of the performance for place of articulation without assuming the context. Table-8.2 shows the performance obtained using such an approach. The parameters used in this study are the same parameters of the transition region as was used in the study in chapter 7.

The places of articulation for Hindi consonants are grouped depending on the manner of the consonant. The stop and nasal consonants occur in five distinct places of articulation, the semivowels in four and the fricatives in four other places of articulation. Therefore, when taking all the cases of CV utterances, a larger set of places are used to differentiate them. Since there is commonality in certain of the places of articulation in the three groups - stop and nasals, semivowels, and, fricatives - a total of eight distinct places have been identified to cover all the three groups. The results of recognition performance shown in Table-8.2 are with respect to these eight places. It is seen from this table that the performance is only about 50% when the manner and vowel contexts are not used. This value is distinctly lower when compared to that obtained when the context was used (about 90% from Table-8.1b).

Since the number of parameters for characterizing the place of articulation is large (ten each from five consecutive overlapping frames), we have reduced the number of parameters to eighteen by using only six of these parameters per frame from three alternate frames. The results for place classification using this reduced parameter set is shown in Table-8.3. The difference in performance when compared to that in Table-8.2 is not significant, although the value is lower in the reduced parameter case. The reduced parameter set is used subsequently to reduce the complexity of input size.

**Table-8.2 :** Classification of place of articulation using all the five transition frames with ten parameters per frame

The network structure is (50, 30, 30, 8)

TEST	REFERENCE							
	Glottal	Velar	Palatal	Retroflex	Alveolar(1)	Alveolar(2)	Dental	Bilabial
Glottal	80.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0
Velar	0.0	72.5	10.0	5.0	5.0	0.0	0.0	7.5
Palatal	4.5	3.5	58.0	14.5	1.0	4.5	11.0	3.0
Retroflex	0.0	9.5	4.0	41.5	5.5	8.0	25.0	6.5
Alveolar(1)	10.0	0.0	0.0	10.0	55.0	0.0	20.0	5.0
Alveolar(2)	0.0	10.0	20.0	10.0	0.0	50.0	0.0	10.0
Dental	0.0	15.0	7.5	15.0	7.5	7.5	47.5	0.0
Bilabial	6.5	20.0	5.0	13.0	1.5	2.5	12.0	39.0
Overall								51.0

**Table-8.3 :** Classification of place of articulation using three transition frames and six parameters per frame

The network structure is (18, 30, 30, 8)

TEST	REFERENCE							
	Glottal	Velar	Palatal	Retroflex	Alveolar(1)	Alveolar(2)	Dental	Bilabial
Glottal	60.0	0.0	10.0	0.0	0.0	0.0	10.0	20.0
Velar	10.0	17.5	20.0	17.5	0.0	5.0	7.5	22.5
Palatal	3.3	5.6	53.3	20.0	1.1	4.5	6.7	5.6
Retroflex	2.7	9.3	9.3	38.7	8.0	5.3	17.3	9.3
Alveolar(1)	10.0	0.0	0.0	10.0	45.0	0.0	10.0	25.0
Alveolar(2)	0.0	0.0	20.0	20.0	0.0	20.0	20.0	20.0
Dental	0.0	10.0	0.0	17.5	2.5	12.5	40.0	17.5
Bilabial	5.3	6.7	8.0	2.7	2.7	5.3	9.3	60.0
Overall								45.0

## 8.2 CLASSIFICATION STUDIES OF EACH REGION USING THE PARAMETERS OF TWO ADJACENT REGIONS

The aim of the present study is to see the possibility of obtaining improved discrimination performance for the different regions by including the context of the other regions. Therefore in the following study the adjacent regions' parameters (and therefore the context) of the utterance are included in the input to the classifier. These were studied for each of the manner, place and vowel types. Since the number of inputs to the network become large when the two regions' parameters are both included, the actual parameters used in this study is reduced. As discussed in the previous section, the parameters representing the place of articulation are reduced by taking only six parameters per frame and using only three frames when used for the context. The manner region parameters are reduced to eight (parameter numbers 1,3,4,7,8,14,15,19 in Table-6.4 in chapter 6) from the original nineteen when used for the context. The vowel region was represented with all the ten parameters used previously.

Table-8.4 shows the performance realized for the manner classification using both the consonantal region parameters (19) for the manner and the transition region parameters (18) for the place of articulation. It is seen from this table that there is no significant improvement (compared to the entry in Table-8.1a) in the manner classification results when the parameters representing the place of articulation are included. This is probably because the large variability of the transition region parameters may not be aiding the manner classification.

Similar study for the determination of the place of articulation using the adjacent regions' parameters was carried out and the results of this study are shown in Tables 8.5 and 8.6. Since the transition region is in between the consonantal region and the vowel region in any utterance, two separate studies were carried out. In the first study (results shown in Table-8.5) the place of

**Table-8.4 : Manner Classification using both manner region parameters (19) and transition region parameters (18)**

**The network structure is (37, 30, 30, 6)**

TEST	REFERENCE					
	VNG	SMV	VAS	ASP	RLT	FRC
VNG	49.3	6.7	0.0	2.7	28.0	13.3
SMV	18.3	51.7	1.7	0.0	18.3	10.0
VAS	4.0	4.0	86.7	5.3	0.0	0.0
ASP	2.7	0.0	5.3	89.3	0.0	2.7
RLT	13.3	1.3	0.0	6.7	57.3	21.3
FRC	6.7	3.3	0.0	1.7	36.7	51.7
					OVERALL	65.2

**Table-8.5 :** Classification of place of articulation using both transition region parameters (8) and manner parameters (18)

The network structure is (26, 30, 30, 8)

TEST	REFERENCE							
	Glottal	Velar	Palatal	Retroflex	Alveolar(1)	Alveolar(2)	Dental	Bilabial
Glottal	50.0	10.0	0.0	0.0	0.0	20.0	0.0	20.0
Velar	5.0	50.0	15.0	7.5	0.0	0.0	5.0	17.5
Palatal	1.1	13.3	66.7	8.9	1.1	2.2	5.6	1.1
Retroflex	1.3	8.0	13.3	36.0	10.7	6.7	12.0	12.0
Alveolar(1)	5.0	0.0	5.0	10.0	55.0	0.0	20.0	5.0
Alveolar(2)	0.0	20.0	10.0	20.0	0.0	30.0	10.0	10.0
Dental	0.0	0.0	5.0	22.5	7.5	2.5	52.5	10.0
Bilabial	1.3	12.0	2.7	13.3	2.7	2.7	13.3	52.0
OVERALL								51.7

**Table-8.6 :** Classification of place of articulation using both transition region parameters (18) and vowel region parameters (10)

The network structure is (28, 30, 30, 8)

TEST	REFERENCE							
	Glottal	Velar	Palatal	Retroflex	Alveolar(1)	Alveolar(2)	Dental	Bilabial
Glottal	30.0	0.0	0.0	0.0	10.0	0.0	20.0	40.0
Velar	7.5	45.0	7.5	7.5	5.0	2.5	10.0	15.0
Palatal	1.1	6.7	67.8	10.0	0.0	1.1	13.3	0.0
Retroflex	1.3	4.0	9.33	37.3	4.0	9.3	21.3	13.3
Alveolar(1)	0.0	0.0	0.0	5.0	60.0	5.0	20.3	10.0
Alveolar(2)	0.0	0.0	10.0	30.0	0.0	30.0	0.0	30.0
Dental	0.0	5.0	2.5	25.0	2.5	0.0	57.5	7.5
Bilabial	5.3	10.7	1.3	5.3	1.3	1.3	13.3	61.3
Overall								53.9

articulation is determined using the reduced manner parameters (8) in addition to the transition region parameters (18). In the second study, the classification of the place of articulation was carried out using vowel region parameters (10) along with the transition region parameters (18) (results shown in Table-8.6). The results in both the studies show no marked improvement in performance to that indicated in Table-8.3. This shows that the use of neighboring regions' context in this way (using the parameters of the neighboring region) does not seem to improve the recognition performance for the place of articulation.

Study for the determination of the vowel type were also carried out with parameters from the adjacent region, namely, the transition region. The study was carried out with transition region parameters (18) and vowel region parameters (10). The results of this study is shown in Table-8.7. It is seen that there is improvement in performance when compared to Table-8.1~.

To see if the use of parameters from all the three regions in the utterance can help to improve the performance, a set of studies were carried out using all parameters from all the regions. This is discussed in the next section.

### **8.3 CLASSIFICATION STUDIES OF EACH REGION USING PARAMETERS OF ALL THREE REGIONS**

Three different studies were carried out using the parameters from all the three regions. These studies used the same parameters as input to the classifier, but the training was carried out by using different network outputs in each case. In the first study, the outputs represent the consonant manner. Hence in this case the network has six outputs. The result of this study is shown in Table-8.8. The results show only a marginal difference when compared to the results discussed in the previous section. This suggests that performance in manner classification

**Table-8.7 : Vowel classification using both transition region parameters (18) and vowel region parameters (10)**

**The network structure is (28, 30,30,5)**

TEST	REFERENCE				
	[I]	[e]	[^]	[o]	[U]
[I]	96.4	3.6	0.0	0.0	0.0
[e]	0.0	100.0	0.0	0.0	0.0
[^]	0.0	0.0	100.0	0.0	0.0
[o]	0.0	0.0	0.0	100.0	0.0
[U]	0.0	0.0	0.0	7.1	92.9
				<b>OVERALL</b>	<b>97.8</b>

can not be improved significantly by using the context of neighboring region in the form of additional parameters in input.

A second study was carried out by using the classifier's output to represent the place of **articulation**. All the three regions' parameters were used as input. It is seen from the results (shown in **Table- 8.9**) that no significant improvement in performance was **obtainable** for the place case as **well**. It maintains the same low level (about 50%) of recognition accuracy.

Studies to determine the performance of the vowel **classification** using **all** three region parameters as input and only vowel **classes** as output were carried out (see **Table-8.10**). The same good recognition was **realized** for the vowel case irrespective of the context.

The third study using **all** the three regions' parameters (8 for manner, 18 for **place** and 10 for **vowel** region) were used as input, and the number of output nodes were increased to 19 (manner **SIX**, place **EIGHT** and vowel **FIVE**) so that **all** the three region **classes** were accommodated. The outputs were grouped into three sets, one set corresponding to each region type. The training was carried out by providing target activation with more than one output being active for each utterance. The results of this study are shown **separately** in **Table-8.11** for the three different regions of the utterance.

Table-8.11 c indicates the performance for the **classification** of the **vowel** type of the utterance. It is noticed that **vowel** recognition performance is very high (97.5%).

**Table-8.11a** shows the performance for the **classification** of consonant manner of the utterance. It is noticed that the performance for this part has also shown some significant improvement (73.5%) compared to the previous studies.



**Table-8.8 :** Manner Classification using all three regions' parameters

The network structure is (47, 30, 30, 6)

TEST	REFERENCE					
	VNG	SMV	VAS	ASP	RLT	FRC
VNG	61.3	9.3	1.3	1.3	22.7	4.0
SMV	20.0	60.0	1.7	0.0	13.3	5.0
VAS	4.0	2.7	88.0	5.3	0.0	0.0
ASP	5.3	0.0	8.0	82.7	2.7	1.3
RLT	14.7	8.0	2.7	0.0	64.0	10.7
FRC	6.7	3.3	0.0	0.0	31.7	58.3
OVERALL						69.8

**Table-8.9 :** Vowel Classification using all three regions' parameters

The network structure is (36, 30, 30, 5)

TEST	REFERENCE				
	[I]	[e]	^	[o]	[U]
[I]	98.2	1.8	0.0	0.0	0.0
[e]	0.0	100.0	0.0	0.0	0.0
^	0.0	0.0	98.2	1.8	0.0
[o]	0.0	0.0	0.0	100.0	0.0
[U]	0.0	0.0	0.0	7.1	92.5
OVERALL					97.8

**Table-8.10 :**Classification of place of articulation using all three regions' parameters

The network structure is (36, 30, 30, 8)

TEST	REFERENCE							
	Glottal	Velar	Palatal	Retroflex	Alveolar1	Alveolar2	Dental	Bilabial
Glottal	40.0	10.0	0.0	10.0	30.0	0.0	0.0	10.0
Velar	2.5	50.0	5.0	22.5	0.0	0.0	10.0	10.0
Palatal	1.1	7.8	73.3	12.2	1.1	1.1	2.2	1.1
Retroflex	2.7	2.7	9.3	48.0	2.7	4.0	26.7	4.0
Alveolar1	5.0	0.0	0.0	10.0	60.0	5.0	20.0	0.0
Alveolar2	0.0	10.0	10.0	0.0	10.0	60.0	0.0	10.0
Dental	0.0	0.0	5.0	20.0	5.0	2.5	60.0	7.5
Bilabial	2.7	16.0	2.7	10.7	1.3	2.7	13.3	50.7
OVERALL								57.2

Table-8.11 :Overall Classification using all three regions' parameters (8+18+10)  
with a single network having all the three types of outputs (6+8+5)

The network structure is (36, 30, 30, 19)

**a) Manner (6 outputs)**

TEST	REFERENCE					
	VNG	SMV	VAS	ASP	RLT	FRC
VNG	76.0	20	0.0	2.0	18.0	2.0
SMV	0.0	82.5	2.5	0.0	2.5	12.5
VAS	0.0	8.0	78.0	14.0	0.0	0.0
ASP	2	20	18.0	78.0	0.0	0.0
RLT	12.0	20	2.0	0.0	72.0	12.0
FRC	0.0	16.7	4.8	7.1	19.1	52.4
OVERALL						73.5

**b) Place of articulation (8 outputs)**

TEST	REFERENCE							
	Glottal	Velar	Palatal	Retroflex	Alveolar1	Alveolar2	Dental	Bilabial
Glottal	36.4	9.1	0.0	0.0	27.3	0.0	18.2	9.1
Velar	0.0	47.5	5.0	22.5	2.5	2.5	10.0	10.0
Palatal	4.9	6.6	62.3	11.5	1.64	1.0	5.0	1.0
Retroflex	0.0	10.0	12.0	54.0	6.0	2.0	10.0	6.0
Alveolar1	1.0	5.0	30.0	5.0	40.0	10.0	0.0	0.0
Alveolar2	20.0	10.0	20.0	0.0	10.0	30.0	10.0	0.0
Dental	2.0	8.0	8.0	26.0	6.0	4.0	32.0	14.0
Bilabial	0.0	0.0	5.0	12.5	2.5	0.0	17.5	62.5
OVERALL								49.6

**c) Vowel (5 outputs)**

TEST	REFERENCE				
	[i]	[e]	[^]	[o]	[U]
[i]	98.2	1.8	0.0	0.0	0.0
[e]	0.0	100.0	0.0	0.0	0.0
[^]	0.0	0.0	98.2	1.8	0.0
[o]	0.0	0.0	0.0	100.0	0.0
[U]	1.8	0.0	0.0	7.2	91.0
OVERALL					97.5

It is to be noted that this improvement is obtained even after a significant reduction in the number of parameters used for the representation of the consonant manner region.

Table-8.11b shows the performance for the place of articulation. It is seen that the performance has not improved for this case. The parameters in the neighboring region do not seem to help to improve the performance for this case. The best approach seems to be the use of independent classifiers for each consonant manner and vowel class combination, as was done in chapter 7. But such an approach implies that the consonant manner and vowel class of the utterance are known apriori (or should be known before hand). This implies that one needs to view the problem in a hierarchical model. We shall discuss this in the next section.

#### **8.4 A HIERARCHICAL MODEL FOR THE RECOGNITION OF CV UTTERANCES**

By the very nature of the problem, the recognition of the place of articulation of a CV utterance with a single network taking all possible contexts is not expected to be high. It is only by considering a smaller subset that the confusability can be expected to be reduced. Therefore it is only within those cases of place of articulation which have the sama consonant manner and vowel context that good recognition is achievable. Fortunately, the vowel class of the utterance can be recognized with good accuracy as indicated in our studies. The manner of the consonant can also be determined but with a lower of performance. Even then, the errors in the determination of vowel and manner are most often to the nearby classes as indicated in the previous chapters. Since the consonant manner determination shows improvement when the parametric context of the different regions in the utterance are also used, this input can be used to obtain improved manner determination accuracy. Since parametric context of the different regions does not call for prior knowledge of the region type, it is possible to

provide this input in advance. Therefore, it is possible to obtain the vowel type and consonant manner of the utterance as a first step. The known context of the vowel type and consonant manner can subsequently be used to select the appropriate classifier for the determination of place of articulation. Thus by carrying out recognition in a hierarchical manner it is possible to increase the recognition performance of **CV** utterances. The top level in the hierarchy addresses the determination of the significant instants in the utterance. With this input the three different regions of a **CV** utterance are determined. At the next level a region specific analyses is carried out to obtain the parameters in different regions of the utterance. The parameters thus obtained are used to determine the vowel class and the consonant manner of the utterance. The results of this second level determine the consonant group and the vowel group to which the utterance belongs. With this input, in the third level of the hierarchy, the specific classifier is chosen for determining the place of articulation of the consonant. The third level is called for mainly to perform selection of the correct classifier for determining the place of articulation.

It is possible to estimate to a first approximation the performance realizable using this hierarchical approach based on studies carried out so far. This estimate gives only a lower bound on the performance, since errors in different parts may also occur simultaneously in the same utterance. Therefore, based on the performance realized for the different parts, namely, about 97% for the vowel classification, about 75% for the manner classification and 90% for the place of articulation in context, the worst case estimate for the overall performance is approximately 65%, being the product of the three performance figures. In the light of the fact that the number of **CV** utterance classes are large (140) and that they are confusable, the estimated performance implies that this performance can be considered as good. Since errors are mostly to the nearby classes for each individual region, overall error will be even lower than this estimate.

The estimated recognition for the best choice is 65%. If we include cases of region classifiers' output wherein the best two choices have the correct class in them then the recognition performance will be significantly better.

## **8.5 SUMMARY**

From the above studies carried out it is seen that recognition of the confusable set of isolated CV utterances can be achieved with good performance by giving proper attention to each distinct part of the utterance. It is possible to obtain improved performance by the use of a hierarchical approach. The proposed hierarchical approach implies that the more dominant features which are more easily discriminable need to be carried out first. The context of these dominant features is used in the later stage to obtain better discrimination of the difficult features.

## **SUMMARY AND CONCLUSIONS**

In this thesis we have studied the issues in the recognition of isolated utterances of Hindi characters. We have addressed signal processing issues dictated by the nature of the signal and have proposed an approach to realize recognition of CV utterances. We shall now summarize the work carried out highlighting the major aspects of the study.

Acoustic-phonetic knowledge of the CV utterances of Hindi was used as the basis for evolving an approach for recognition. The CV utterances were grouped into broad classes based on the features that characterize the classes. Relevant regions in a CV utterance that carry important information required to discriminate different classes were identified. The region prior to the vowel onset instant in the utterance was identified as the region that represents the consonant manner. The region after vowel onset in the utterance was identified as the relevant part for representing the vowel characteristics of the utterance. The articulatory release instant was identified as an important event signifying the start of the transition region. Then the issues involved in the recognition of CV utterances were identified as follows :

1. Identification of the three significant instants - start of the utterance, the articulatory release instant and the vowel onset instant.
2. Recognition of the vowel class from the signal in the vowel region of the utterance.

3. Recognition of the consonant manner from the signal in the region before the vowel onset.
4. Recognition of the place of articulation knowing the consonant manner and the vowel context by analyzing the transition region immediately following the articulatory release instant.
5. Development of an overall recognition strategy.

By identifying the issues involved in classification of different regions of **CV** utterances, the main issue in the recognition of **CV** utterances was identified as processing of the signal in the regions appropriately to detect the required features. The time normalization that is normally required to overcome the durational variability in utterances is no longer required in this approach. The focus was thus on the issues in detecting the required features in different regions of the utterance.

Having identified the five specific issues, each issue was taken up as an independent study. Since the basis for the independent study was the analysis of the relevant region in each **CV** utterance, each study required as its input the boundaries of the corresponding region. These were obtained through manual analysis of the utterance data used in the studies.

Recognition of the significant instants was the first issue to be addressed. These instants could be identified with the associated source of excitation and signal processing methods using the minimum phase property of the signal. This approach enabled us to pick up the significant instants with very good time resolution. Postprocessing was used to discard some spurious epochs. This approach was used mainly to aid in the manual identification of region boundaries. Since the computational burden due to this approach was very high, parameters based on block data processing were suggested in the study for detection of the

instants. The choice of parameters was determined based on the region characteristics on either side of the instant. A neural network classifier was used for learning the features in the parameters which characterize the change in the signal at the significant instant. Good performance was achieved in the identification of instants by these methods.

The next study addressed some issues in the identification of the vowel class. Although spectral features characterize the vowel region, the more direct clues to the vowel class were considered. In particular, area coefficients derived from LPCs were used to describe the vocal tract shape in the vowel region. The variability in the area function called for normalization before they could be used. Weighted cepstral parameters were also used to obtain comparative performance. Neural network classifiers were developed for vowel discrimination using different parametric representations as input. The performance of vowel discrimination with weighted cepstral coefficients was good. The area coefficients exhibited lower performance but still comparatively good considering the nature of variability in them.

In the study on recognition of the consonant manner, the region from the start of utterance till vowel onset was identified as the relevant part for analysis. Gross parameters in this region were used to represent the consonant manner. Trend and spectral distance parameters were also included to represent the regions' features better. Neural network classifiers were developed to discriminate between the different manners of consonant. The discrimination performance was better when trend parameters were also included. Studies were carried out to see if improvement in performance can be obtained by considering the vowel context. The results indicated about the same performance.

In the study on discrimination of place of articulation, the transition region immediately following the articulatory release instant was identified as the relevant



part for analysis. Several consecutive frames in a fixed duration of the transition region were used to characterize the dynamic nature of the articulators of this region. Area coefficients were again used as in the case of vowel classification, with cepstral coefficients for comparison. As the nature of the transition region depended on the consonant manner and vowel, different neural network classifiers were developed for discrimination of the place of articulation in different contexts. Performance comparison for different parametric representations was carried out.

Finally, an approach to the overall CV utterance recognition has been proposed. The aim of this scheme was to minimize the errors in the overall recognition by using the embedded context that the manner, place and vowel correspond to the same utterance. Using the results of the performance based on different schemes for using the context of the neighbour region in a CV utterance, a method was evolved to realise the overall CV utterance recognition. The estimated performance of such a system is about 65% for all the 140 classes.

Better performance is possible if this isolated utterance is in the context of words and sentences. In this context, higher level language rules can be made use of to disambiguate multiple choices from the output of the classifier.

For a recognition system to be usable, it is necessary to explore parameters and features which are speaker independent. It is for this reason that the use of area coefficients were explored. Area coefficients though they are appealing because of their closeness to articulatory features, they do not perform as expected. Area coefficients as derived from LP analysis show large variability. Their poor performance is attributable to this variability. The variations in the area coefficients due to the analysis method is probably masking the invariant articulatory features such as back, central, front, high, mid, low, rounded and

unrounded. Different analysis methods which can bring out these articulatory features are needed to obtain speaker independent recognition.

It is seen from the discussion in this thesis, that the nature of speech data and the characterizing features for discrimination have fuzzy description. Future directions to this work can exploit this fuzzy nature of the input and output by designing the classifier accordingly.

## REFERENCES

- [Ananthapadmanabha 1975] T.V.Ananthapadmanabha and B.Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.23, pp.562-570, 1975.
- [Ananthapadmanabha 1979] T.V. Ananthapadmanabha and B.Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.27, pp.309-319, 1979.
- [Atal 1971] B.S.Atal and S.L.Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol.50, pp.637-655, 1971.
- [Atal 1976] B.S.Atal and L.R.Rabiner, "A pattern recognition approach to voiced/unvoiced/silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.24, pp.201-212, 1976.
- [Broad 1975] D.J.Broad and J.E.Shoup, "Concepts in speech recognition," in *Speech Recognition*, D.Raj Reddy, Ed. New York: Academic Press, pp.243- 274, 1975.
- [Burr 1988] D.J.Burr, "Experiments on neural net recognition of spoken and written digits," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.36, pp.1162- 1168, 1988.
- [Carlson 1979] R.B.Carlson, Granstrom and D. Klatt, "Some notes on the perception of temporal patterns in speech," in *Frontiers in Speech*

*Communication Research*, B.Lindholm and S. Ohman, Eds. Academic Press, pp.233-243, 1979.

[Catford 1988] J.C.Catford, *Practical Introduction to Phonetics*, Oxford: Oxford University Press, 1988.

[Cheng 1989] Y.M.Cheng and D.O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.37, pp.1805-1815, 1989.

[Cole 1990] R.A.Cole, M.Fanty, Y.Muthusamy and M.Gopalakrishnan, "Speaker independent recognition of spoken English letters," *Proc. IEEE Int. Joint Conf. on Neural Networks*, San Diego, 1990.

[Cole 1991] R.A.Cole, M.Fanty, M.Gopalakrishnan and R.Janssen, "Speaker independent name retrieval from spellings using a database of 50,000 names," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, Toronto, 1991, pp.325-328.

[Davis 1980] S.B.Davis and P.Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, ~01.28, pp.357-366, 1980.

[Demichelis 1989] P.Demichelis, L.Fissore, P.Laface, G.Micca and E.Piccolo, "On the use of neural networks for speaker independent isolated word recognition," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, Glasgow, Scotland, pp.314-317, 1989.

[Fant 1970] G. Fant, Acoustic Theory of Speech Production, 2nd ed., The Hague: Mouton, 1970.

[Fant 1973] G. Fant, Speech Sounds and Features, Cambridge: MIT press, 1973.

[Flanagan 1972] J.L. Flanagan, Speech Analysis Synthesis and Perception, 2nd ed., New York: Springer-Verlag, 1972.

[Freisleben 1992] B.Freisleben and C.A.Bohn, "Speaker independent recognition of word with back propagation networks," in Artificial Neural Nets and Genetic Algorithms, R.F.Albrecht, C.P.Reeves and N.C.Steele, Eds. New York: Springer Verlag, pp.243-248, 1992.

[Fry 1979] D.B.Fry, The Physics of Speech, Cambridge: Cambridge University Press, 1979.

[Ganesan 1975] S.N. Ganesan, Contrastive Grammar of Hindi and Tamil, Madras: University of Madras, 1975.

[Gray 1974] A.H.Gray and J.D. Markel, "A spectral flatness measure for studying the auto-correlation method of linear prediction of speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.24, pp.207-217, 1974.

[Harrington 1988] J. Harrington, "Automatic recognition of English consonants, in Aspects of Speech Technology, M.A.Jack and J.Laver, Eds. Edinburgh: Edinburgh University Press, pp.69-143, 1988.

- [Hermansky 1986] H.Hermansky, K.Tsuga, S.Makino and H.Wakita,"Perceptually based processing in automatic speech recognition," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp.1368-1371, 1986.
- [Hush 1993] D.R.Hush and B.G.Horne, "Progress in supervised neural networks," *IEEE Signal Processing Magazine*, 9- 39, Jan. 1993.
- [Jayant 1984] N.S.Jayant and P.Noll, *Digital Coding of Waveforms - Principles and Applications to Speech and Video*, New Jersey: Prentice-Hall Inc., 1984.
- [Kewley-Port 1982] D.Kewley-Port, "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," *J. Acoust. Soc. Am.*, vol.72, pp.379-389, 1982.
- [Klein 1970] W.Klein,R.Plomp and L.C.W.Pols, "Vowel spectra, vowel spaces and vowel identification," *J. Acoust. Soc. Am.*, vol.48, no.4(Part 2), pp.999-1009, 1970.
- [Kuhn 1975] G.M. Kuhn, 'On the front-cavity resonance and its possible role in speech perception", *J. Acoust. Soc. Am.*, vol.58, vo.2, pp.428-433, 1975.
- [Lang 1992] K.J.Lang, A.Waibel and G.E.Hinton, "A time-delay neural network architecture for isolated word recognition," in *Artificial Neural Nets - Paradigms, Applications and Hardware Implementations*, E.Sanchez-Sinencio and C.Lau, Eds. IEEE Press, pp.388- 408, 1992.
- [Ladefoged 1975] B. Ladefoged, *A Course in Phonetics*, New York: Harcourt, Brace Jovanovich, 1975.

- [Lea 1980] W.A. Lea, Trends in Speech Recognition, New York: Prentice Hall, 1980.
- [Lee 1993] L.Lee, C.Tseng, H.Gu, F.Liu, C.Chang, Y.Lin, Y.Lee, S.L.Tu, S.H.Hsieh and C.Chen, "Golden Mandarin(I) - A real-time Mandarin speech dictation machine for Chinese language with very large vocabulary," *IEEE Trans. Speech, Audio Processing*, vol.1, no.2, pp.158-179, 1993.
- [Lennig 1984] M. Lennig and P. Brassard, "Machine-readable phonetic alphabet for English and French," *Speech Communication*, vol.3, pp.166, 1984.
- [Lieberman 1956] A.M.Lieberman, P.C.Delattre, F.S.Cooper and L.J.Gernstrom, "Tempo of frequency changes as a cue for distinguishig classes of speech sounds," *J Acoust. Soc. Am.*, vol.52, pp.127-137, 1956.
- [Lienard 1984] J.S.Lienard and F.K.Soong, "On the use of transient information in speech recognition," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, pp.17.3.1-17.3.4, 1984.
- [Lipman 1987] Richard Lippman, "An Introduction to computing with neural networks", *IEEE ASSP Magazine*, vol.4, pp.4-22, Sep. 1987.
- [Markel 1973] J.D. Markel, "The SIFT algorithm for fundamental frequency estimation", *IEEE Trans. Audio and Electro Acoustics*, AU-21, no.3, pp.149-153, 1973.
- [Markel 1974] J.D.Markel and A.H.Gray, Jr., *Linear Prediction of Speech*, New York: Springer Verlag, 1974.

- [Murthy 1991a] H.A. Murthy and B. Yegnanarayana, "Formant extractions from group delay functions," *Speech Communication*, Vol.10, pp.209-221, 1991.
- [Murthy 1991b] H.A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," *Signal Processing*, vol.22, No.3, pp.259-267, 1991.
- [Myers 1981] C.S.Myers and L.R.Rabiner, "A comparative study of several DTW algorithms for connected word recognition," *Bell System Technical Journal*, vol. 60, no. 7, 1981, pp. 1389-1409.
- [Nathan 1991] K.S.Nathan, Y. Lee and H.Silverman, "A time varying analysis method for Rapid transitions in speech," *IEEE Trans. Signal Processing*, vol.39, no.4, 1991.
- [Niederjohn 1985] R.J.Niederjohn and M.Lahat, "A zero-crossing consistency method for tracking of voiced speech in high noise levels," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.33, pp.349-355, 1985.
- [Parthasarathy 1987] S. Parthasarathy and Donald W. Tufts, "Excitation-synchronous modeling of voiced speech," *IEEE Trans. on Acoust. Speech, Signal Processing*, vol.35, no.9, pp.1241-1249, 1987.
- [Rabiner 1975] L.R.Rabiner and Sambur, "An algorithm for determining the end points of isolated utterances," *Bell System Technical Journal*, vol.54, pp.297-316, 1975.



- [Rabiner 1978a] L.R.Rabiner, "On creating reference templates for speaker independent recognition of isolated words," IEEE Trans. Acoust., Speech, Signal Processing, vol. 26, pp.34-42, 1978.
- [Rabiner 1978b] L.R.Rabiner and R.W.Schaffer, Digital Processing of Speech Signals, Prentice Hall, New Jersey, 1978.
- [Rabiner 1981] L.R.Rabiner and S.E.Levinson, "Isolated and connected word recognition - Theory and Selected Applications," IEEE Trans. Communications, COM-29, pp. 621-659, 1981.
- [Rabiner 1989] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol.77, no.2, pp.257-286, 1989.
- [Rabiner 1993] L.R.Rabiner and B.Juang, Fundamentals of Speech Recognition, New Jersey: Prentice-Hall, pp.242-319, 1993.
- [Raman 1985] S.Raman, Speech recognition of Hindi stop consonants, Ph.D. Thesis, I. I. T. Madras, India, 1985.
- [Ross 1974] M.J.Ross, H.L.Schaffer, A.Cohen, R.Freudberg and H.J.Mandey, "Average magnitude difference function pitch extractor," IEEE Trans. Acoust., Speech, Signal Processing, vol.22, pp.353-362, 1974.
- [Sakoe 1978] H.Sakoe and S.Chiba, "Dynamic programming optimization for spoken word recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol.26, pp.43-49, 1978.

- [Shirai 1991] K. Shirai and T. Kobayashi, "Estimation of articulatory motion using neural networks," *J. Phonetics*, vol.19, pp.379-385, 1991.
- [Smits 1992] R.L.H.M. Smits and B. Yegnanarayana, 'Determination of instants of significant excitation in speech using group delay functions,' Manuscript no.886, Institute for Perception Research, Eindhoven, 1992.
- [Sundar 1987] R. Sundar, S. Raman and B. Yegnanarayana, "Studies on speech recognition of Hindi stop consonants," *Proc. Euro. Conf. Speech Technology*, Edinburgh, pp.95-98, Sep. 1987.
- [Viswanathan 1975] R.Viswanathan and J.Makhoul, "Quantization properties of transmission parameters in linear prediction systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, pp.309-321, 1975.
- [Wakita 1974] H.Wakita, "Estimation of vocal tract length from acoustic data," *J. Acoust. Soc. Am.*, vol. 55, supplement J-7, 1974.
- [Wakita 1975] H.Wakita and Augustine H.Gray Jr. "Numerical determination of the lip impedance and vocal tract area functions," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.23, 1975, pp.574-580.
- [Wakita 1977] H. Wakita,"Normalization of vowels by vocal tract length and its applications to vowel identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.25, pp.183-190, 1977.
- [Yegnanarayana 1979] B. Yegnanarayana and D. Raj Reddy, "A distance measure based on the derivative of linear prediction phase spectrum," *Proc. Int. Conf. Speech, Signal Procesing*, pp.744-747, 1979.

[Yegnanarayana 1981] B. Yegnanarayana, "Speech analysis by pole- zero decomposition of short-time spectra," *Signal Processing*, vol.3, pp.5-17, 1981.

[Yegnanarayana 1984] B. Yegnanarayana and T. Sreekumar, "Signal dependent matching for isolated word speech recognition system," *Signal Processing*, vol.7, pp.161-173, 1984.

[Yegnanarayana 1986] B. Yegnanarayana, S. Raman and R. Sundar, "Signal dependent analysis for speech recognition," *Proc. IEEE Int. Conf. Speech Input/Output Techniques and Applications*, London, pp.31-36, Mar. 1986.

[Yegnanarayana 1991] B. Yegnanarayana and R. Sundar, "Signal processing issues in realizing voice input to computers," *Asia- Pacific Engineering Journal (Part A)*, vol.1, no.2, pp.197-217, 1991.

## **LIST OF PUBLICATIONS**

1. [Yegnanarayana 1991] B. Yegnanarayana and R. Sundar, "Signal processing issues in realizing voice input to computers," Asia-Pacific Engineering Journal (Part A), Vol. 1, no. 2 1991, pp.197-217.
2. [Sundar 1987] R. Sundar, S. Raman and B. Yegnanarayana, "Studies on speech recognition of Hindi stop consonants," Proc. of Euro. Conf. on Speech Technology, Edinburgh, Sep. 1987, pp. 95-98.
3. [Yegnanarayana 1986] B. Yegnanarayana, S. Raman and R. Sundar, "Signal dependent analysis for speech recognition," Proc. IEEE Int. Conf. Speech Input/Output Techniques and Applications, London, March 1986, pp. 31-36.