To

the memory of

my father

Sri C.N.RADHAKRISHNAIAH CHETTY

THESIS CERTIFICATE

This is to certify that the thesis entitled Neural Network Models for Recognition of Stop Consonant-Vowel (SCV) Segments in Continuous Speech submitted by **C.Chandra** Sekhar to the Indian Institute of Technology, Madras for the award of the degree of Doctor of Philosophy is a bona fide record of research work carried out by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Madras **600 036** (B.Yegnanarayana)

Date: April [illegible], 1996

ACKNOWLEDGMENTS

'Research is to SEE what everybody else has seen, and to THINK what nobody else has thought'

- Anonymous

The latter part of research as per the above quote is considered to be an order of magnitude more difficult than the former part. Long term success and enjoyment in research lies in constantly seeking for and pursuing unconventional solutions. 1 have been fortunate to be associated with Prof.B.Yegnanarayana who has always emphasized the need for and the usefulness of unconventional thinking in generating research ideas. 1 have been immensely benefitted in many ways from the discussions and interaction I have had with him throughout this work. 1 am extremely grateful to him for his inspiration, encouragement and guidance.

'Feeling gratitude and not expressing it is like wrapping a present and not giving it'

- William Arthur Ward
(Readers Digest, April 1996, pp.186)

I have received the necessary support from the Heads of the Department during this thesis work. I would like to thank Prof.R.Nagarajan, Prof.Kamala Krithivasan and Prof.S.V.Raghavan for their support. I acknowledge the cooperation extended by my faculty colleagues. I would like to specifically mention the help rendered by Dr.R.Sundar and N.Alwar. I would like to thank all my doctoral committee members for their interest and suggestions.

I am grateful to Prof.K.Nagamma Reddy, Osmania University, Hyderabad, from whom I have learnt acoustic-phonetics without which I would have been less comfortable in carrying out this work. I have been greatly benefitted from the discussions I had with Dr.R.Sundar, S.Rajendran, Dr.P.P.Raghu and Dr.Harish M. Chouhan, and I am indebted to all of them. I gratefully acknowledge the keen interest evinced and the help extended by Dr.G.V.Ramana Rao, Dr.P.Eswar and Dr.Hema A. Murthy. I am thankful to the M.Tech. students, S.R.Prakash, Santosh S. Mathews and K.Sakthivel for their assistance in carrying out some preliminary studies.

I would like to thank many colleagues in Speech and Vision Laboratory who helped me in many ways during this research work. I would like to specifically mention S.Rajendran who has always been willing to help me in collection of speech data and also in using many tools for preparation of this thesis. I acknowledge the help that I received from P.Anjaneyulu, N.Sudha and P.Sathyanarayana Murthy.

Many personal friends and relatives have readily extended their moral support. I would like to mention the keen interest shown and the encouragement given by my brother, Dr.C.Sreeramulu Chetty, and my close friend, D.Audikesavulu Naidu. I am extremely thankful to my mother for her affection and care. I acknowledge the cooperation extended by the other members of my family and my in-laws. I cannot forget to mention the interest shown and the help rendered by the cheerful friend, G.Padma.

I am bereft of words to express my appreciation for the understanding and cooperation of my wife, Anitha. Finally, it is the joyful time with our little Abhiroop that provided the much needed relaxation and helped in easing my tensions.

C. Chandra Sekhar

ABSTRACT

This thesis addresses some issues in the recognition of subword units in continuous speech. The main issues addressed are related to handling the large number of Stop Consonant-Vowel (SCV) units and the confusability among these units. To deal with these issues a new feature is presented in this thesis, namely use of the acoustic-phonetic knowledge to improve the classification performance. The knowledge is incorporated in the form of constraints in a constraint satisfaction model. The significant feature of this model is that a collection of even weak evidences could enhance the discriminability of confusable units.

Modular neural networks are considered for developing classifiers for the large number of classes. Separate neural networks (subnets) are trained for subgroups of classes. The performance of the conventional modular networks is poor because classification is performed by assigning the class of the largest value among the outputs of the subnets. We develop a Constraint Satisfaction Model (CSM) in which the outputs of the subnets are combined using the constraints that represent the similarities among the SCV classes. The constraints are derived from the acoustic-phonetic knowledge of the classes and also from the performance of the subnets. The improved performance of the CSM is mainly due to its ability to enhance even the weak evidences and combine the multiple evidences available in the outputs of the subnets based on different grouping criteria. Though the CSM is developed for the classification of isolated utterances of SCVs, the approach can be extended for the classification of SCV segments in continuous speech.

For spotting the SCV segments in continuous speech, an approach based on the detection of Vowel Onset Points (VOPs) and scanning around the VOPs using the classifiers is developed. This approach is shown to be useful in reducing the number

of false alarms, besides reducing the computational complexity significantly. A neural network based method is proposed for the detection of VOPs in continuous speech.

An analysis of the performance of the models for recognition of SCVs has shown that a significant percentage of errors is due to misclassification of the place of articulation of the stop consonants. The place of articulation information is reflected in the formant transitions, and hence suitable methods for extracting and representing the formant transition information are explored.

The methods presented in this thesis suggest ways of dealing with large number of confusable subword units like SCVs, which in turn may lead to the realization of a speech signal-to-symbol transformation module of a speech recognition system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ACON	All-Class-One-Network
ANN	Artificial Neural Network
CSN	Constraint Satisfaction Model
CV	Consonant-Vowel
DHMM	Discrete Hidden Markov Model
FMT	Formant Frequencies and Amplitudes
HMM	Hidden Markov Model
Hz	Hertz
LP	Linear Prediction
MCEP	Mel-scale Weighted Cepstral Coefficients
MLFFNN	Multilayer Feedforward Neural Network
MLP	Multilayer Perceptron
MMI	Maximum Mutual Information
MOA	Manner of Articulation
MS	Multispeaker
ms	milliseconds
OCON	One-Class-One-Network
POA	Place of Articulation
PR	Pitch Region Analysis
SCV	Stop Consonant-Vowel
SD	Speaker Dependent
SI	Speaker Independent
SPC	Spectral Coefficients
TDNN	Time Delay Neural Network
UVA	Unvoiced-Aspirated
UVUA	Unvoiced-Unaspirated
VA	Voiced-Aspirated
VC	Vowel-Consonant
VOP	Vowel Onset Point
VUA	Voiced-Unaspirated
WCEP	LP-derived Weighted Cepstral Coefficients

Chapter 1

INTRODUCTION

1.1 Problem of Continuous Speech Recognition

One of the major research problems in the field of artificial intelligence is to provide natural input/output to a computer. Natural communication can be through either speech or image. In this thesis, we address some issues involved in providing natural input through speech in Indian languages. We are specifically interested in continuous speech recognition where the clearly spoken input speech is converted into a meaningful text. This provides a limited dictation capability to a computer. **A** continuous speech recognition system consists of two major modules, namely, (1) speech signal-to-symbol transformation module and **(2)** symbol-to-text conversion module. The signal-to-symbol transformation module converts the input speech into a sequence of symbols which will be converted into a meaningful text by the symbol-to-text conversion module using lexical, syntactic and semantic knowledge sources. In the present thesis work, we address some issues involved in developing a signal-to-symbol transformation module for vocabulary independent recognition of continuous speech in Indian languages.

Signal processing and pattern matching techniques used for isolated word speech recognition cannot be extended for continuous speech recognition because of the complex variations in the characteristics of basic speech units due to coarticulation, context anti speaking rate in continuous speech. Human beings use language specific knowledge in addition to the knowledge of speech production in identifying different segments of continuous speech [1]. For a computer to perform the same task, it is necessary to endow it with this knowledge. Initially it was proposed to develop the signal-to-symbol transformation module based on expert systems approach using primarily the acoustic-phonetic knowledge. Characters of an Indian language (Hindi) were chosen as symbols and a character spotting approach to develop the signal-to-symbol transformation module was explored [2].

One of the main difficulties in the expert systems approach was the acquisition of knowledge necessary to build the knowledge-base for the expert system. It is difficult to manually derive the rules and refine them. The difficulty is mainly due to the complex nature of the continuous speech. The manifestation of a particular sound unit in the speech signal is different not only for different speakers, but also for the same speaker in different contexts due to coarticulation effects. It is not possible to collect and enumerate all types of variations and incorporate them explicitly in the form of rules. Acquisition of knowledge was one of the main limitations in the expert systems approach.

Attempts were made to develop recognition models that are capable of acquiring knowledge from examples automatically . Statistical models such as Hidden Markov Models (HMMs) have been extensively used for speech recognition [3] [4]. In approaches based on HMMs, a separate model is trained for each symbol to estimate the probability of a given speech segment being generated by that model. The model parameters are estimated from a large number of examples of the utterances of the corresponding symbol. Successful recognition systems using HMMs have been developed for large vocabulary isolated word recognition and connected word recognition in which words are used as basic speech units. The success of HMMs in vocabulary independent continuous speech recognition has been limited mainly because of the poor capability of these models in discrimination of subword units.

Artificial Neural Network (ANN) models have been shown to be suitable for pattern recognition tasks because of their ability to form complex decision surfaces by using discriminatory learning algorithms [5]. ANN models have been extensively used for speech recognition applications [6] and have been shown to give a better classification performance for recognition of subword units such as phonemes and CV syllables [7]. The main limitation of the ANN models is their inability to model temporal sequences. Recent approaches for continuous speech recognition have been based on hybrid models in which ANN models are used for modeling subword units and HMMs are used for modeling words and sentences as concatenations of subword unit models [8]. The systems are trained and evaluated at the level of sentences. These approaches are suitable for vocabulary dependent continuous speech recognition applications. In order to realize a vocabulary independent continuous speech recognition system, it is necessary to develop models that can recognize subword units in continuous speech independent of vocabulary and task.

In this thesis, we address issues in developing neural network models for vocabulary independent recognition of subword units corresponding to CV segments that occur in continuous speech.

1.2 Approaches to Recognition of Subword Units

Automatic segmentation of continuous speech into subword units and labeling the units by classification has been the standard approach adopted for vocabulary independent recognition of subword units in continuous speech. Commonly used subword units are phonemes, syllables, diphones and triphones. This approach requires a robust and speaker independent method for automatic segmentation of continuous speech into regions corresponding to subword units chosen for a system. Because of the variability in the characteristics of subword units in continuous speech, segmentation is not a trivial task. Once segmentation is done, then the recognition performance is dependent on the ability of classification models to correctly label the subword unit regions.

An alternate approach that avoids the need for automatic segmentation is to spot subword units in continuous speech. In spotting approach, classification models developed for subword units are used to scan the continuous speech signal and identify the regions where the corresponding subword units are present. Recognition performance of systems based on spotting approach depends on the ability of classification models to correctly identify the regions belonging to the corresponding subword units and reject all other regions.

In the following sections, we briefly discuss the main issues in classification and spotting of subword units corresponding to CV segments.

1.3 Importance of Recognition of CV Segments

Phonemes, the sounds corresponding to consonants and vowels, are the basic speech units of a language. Consonants cannot be produced in isolation. A consonant is mostly followed or preceded by a vowel to form a CV or VC speech production unit respectively. A consonant can also be followed or preceded by other consonants to form consonant clusters. But, any consonant cluster should be followed or preceded by a vowel. This results in C*V or VC* speech production units. Here C* denotes the presence of one or more consonants in a unit. A meaningful word, in general, can be considered as a sequence of C*V units or as a sequence of VC* units. It is easier to mark, at least manually, the boundaries of a C*V segment in continuous speech signal than that of a VC* segment. Additionally, in many Indian languages, a C*V unit is represented by a single, may be a composite, character making it a convenient general form of speech production units for processing. Consonant-Vowel (CV) units occur on their own as characters (for example, /ti/ in /prativa:d/) and as part of cluster characters (for example, /ti/ in /mu:rti/). Therefore CV units occur with a high frequency in any text.

In continuous speech, the production of a sound is affected mainly by the immediately adjacent sounds leading to coarticulation effects. Significant clues for recognition of consonants are influenced by their adjacent vowels. Because most of the segmental coarticulation effects are also captured in the CVs, we have considered CVs as basic speech units for processing and performing continuous speech recognition. Development of strategies for redognition of CVs in continuous speech with good accuracy is important for realizing a continuous speech-to-text conversion system for Indian languages.

1.4 Frequency of Occurrence of Different Categories of CVs

A study was carried out to determine the frequency of occurrence of different categories of CVs based on the broad manner of production of consonants. For this study, consonants of Hindi are grouped into the following categories : (1) Stop consonants, (2) Affricates, (3) Nasals, (4) Semivowels and (5) Fricatives. The percentage distributions of different categories indicating their frequency of occurrence in a total of about 16000 CVs present in about 800 Hindi sentences collected from five different texts are given in Table-1.1.

It can be noted that nearly 45% of the total number of CVs belong to the category of Stop Consonant-Vowel (SCV) units. Recognition of SCVs is an important and challenging task because of the high frequency of occurrence of SCVs, the large number (160) of SCV classes and confusability amongst several SCV classes. In this thesis, we focus on recognition of SCV segments that occur in continuous speech.

In the following sections, we discuss the important issues in recognition of SCV

Table 1.1: Percentage distribution indicating the frequency of occurrence of different categories of CVs.

Category of CVs	Percentage Distribution
Stop Consonants	44.32
Affricates	05.63
Nasals	13.40
Semivowels	20.21
Fricatives	16.44

segments in continuous speech. We first discuss the issues in classification of SCV segments excised from continuous speech and then discuss the issues in spotting SCV segments.

1.5 Issues in Classification of SCV Segments

1.5.1 Confusability amongst SCV Classes

The similarities in the nature of speech signal for different SCVs are because of similarities in their production mechanism. Production of SCVs consist of all or a subset of the following significant events: closure, burst, aspiration, transition and vowel. The discriminatory clues for any two SCVs are dependent on the events present in their production and the differences in the characteristics of those events. The manifestation of these differences in the acoustic signal depends on the specific speech production features characterizing the events. Each of the stop consonants in Hindi is uniquely described by its manner of articulation and place of articulation. The differences in the manner of articulation of consonants of two different SCVs manifest as signals of different characteristics (voiced/unvoiced) in the closure region and, as presence, weak presence or absence of the aspiration region. The differences in the place of articulation of consonants of two different SCVs manifest as follows: (1) Different characteristics of the spectrum for the signal in the burst region and (2) Different characteristics of the spectrum in the transition region. The differences in the vowels of two different SCVs mainly affect the characteristics of the signal in the vowel and the transition regions.

Difficulty in classification of SCVs is mainly due to the need for fine discrimination amongst many similar classes. The classification performance depends on the ability of parametric representations of SCV segments to capture the discriminatory information and the capability of classification models to form nonlinear complex decision surfaces in the parametric space. The nature of the SCV utterances suggests that any classifier model needs to use the information present in all the regions of an SCV segment to perform classification. It is not possible to perform classification based on the information in only one or two frames.

1.5.2 Characteristics of SCVs in Continuous Speech

Characteristics of subword units in continuous speech are different from the ideal characteristics of the isolated utterances of units. There is an analogy between continuous speech and cursive script. In continuous speech, the characteristics of subword units are affected by the context and speaking rate. In cursive script, the features of individual characters are significantly modified by the context. The inherent characteristics of the individual units are dep‹·ndent on their production mechanism. The variability in the characteristics of units is mainly due to the context in which they occur.

The context in which an SCV is uttered in continuous speech affects its characteristics. The voicing in the closure regions for voiced SCVs may be present weakly or may not be present at all. The burst may be totally absent. The aspiration region supposed to be present for the aspirated SCVs may also have a weak presence or may not be present at all. The duration of the steady region of the vowel in an SCV is severely affected in continuous speech. This is due to the tendency of the speech production mechanism to use minimum effort in production of sounds in continuous speech. Therefore, the discriminatory clues necessary for classification may be partially present or may be completely absent. A system for recognition of SCVs in continuous speech should be able to classify an SCV with the modified characteristics as one of the SCVs that are close to it in terms of production features. Fuzzy logic can be used to handle the variations in the characteristics of the signal due to variations in speakers and due to the context. Fuzzy logic can also be used to indicate the graded presence of the clues, and to assign confidence levels to the hypotheses made by the classifier. Ordering the hypotheses based on the confidence levels and considering more than one hypothesis in labeling SCV segments helps to some extent in handling the variability in the characteristics of SCVs in continuous speech.

1.5.3 Large Number of SCV Classes in Indian Languages

An important issue in recognition of SCVs in Indian languages is the large number of SCV classes. The total number of possible SCVs for Hindi is 160. This number is arrived at by considering the combination of each of the 16 stop consonants (/k/,/kh/,/g/,/gh/,/ṭ/,/ṭh/,/ḍ/,/ḍh/,/t/, /th/,/d/,/dh/,/p/,/ph/,/b/ and /bh/) with 10 vowels (/a/,/a:/,/i/,/i:/,/u/, /u:/,/e/,/ai/,/o/ and /au/). It is observed that the characteristics of an SCV with a consonant and a short vowel (/a/,/i/ or /u/) are more or less same as that of an SCV with the same consonant and long version of the vowel (/a:/,/i:/ or /u:/ respectively). The difference is mainly in their overall durations. Therefore such classes can be combined into a single class. This results in a total number of 112 classes. If one is considering only pure vowels (i.e., if diphthongs /ai/ and /au/ are not being considered), then the number of SCV classes is 80. The recognition system for such a large number of classes should use a suitable classifier architecture which can discriminate all the classes.

1.5.4 Varying Frequency of Occurrence of SCV Classes

It is observed that all the SCVs do not occur with the same frequency. Some SCVs occur very frequently, some less frequently, and others rarely. The frequency of occurrence for different SCV classes in about 800 Hindi sentences is given in Table-1.2. The frequency of occurrence for a class is given as a percentage of the total number (about 7150) of SCVs present in the sentences. It is observed that only 20 SCV classes (out of a total of 80 classes) occur with a frequency greater than the average frequency of occurrence (i.e., 1.25%), and about 20 other classes have a frequency of occurrence that is less than 0.1%.

The varying frequency of occurrence for different classes is a characteristic of any language. This factor needs to be taken into account in evolving the strategies for training and recognition in the design of a classifier. The standard procedure for building the training data sets is to create a database consisting of a large number of sentences collected, and then use a percentage of the totally available data for each class as training data for that class. Because of the different frequencies of occurrence for different classes, the sizes of the totally available data sets for different classes in any database are not uniform. This results in nonuniform sizes of the training data sets for different classes. Because the varying frequency of occurrence of classes is a characteristic of the language, increasing the number of sentences in the database to any extent arbitrarily will not help. One can carefully collect more number of sentences that contain those classes for which the sizes of available data sets are small and try to make the sizes of all classes uniform. This is a difficult task when

Table 1.2: Frequency of occurrence of different SCV classes in Hindi. About 20 classes occur with a frequency greater than the average frequency of occurrence. About 20 other classes occur rarely with a frequency less than 0.1%. This shows that SCV classes have varying frequency of occurrence.

Class	in%	Class	in%	Class	in%	Class	in%
ka	12.68	kha	1.22	ga	3.66	gha	0.48
ṭa	0.85	ṭha	0.32	da	1.26	dha	0.20
ta	6.28	tha	2.64	da	2.64	dha	1.62
pa	7.06	pha	0.39	ba	5.61	bha	1.66
ki	8.65	khi	0.35	gi	0.76	ghi	0.04
ṭi	0.53	ṭhi	0.29	di	0.78	dhi	0.13
ti	3.71	thi	1.54	di	2.06	dhi	0.90
pi	0.67	phi	0.20	bi	0.80	bhi	2.50
ku	0.80	khu	0.15	gu	0.41	ghu	0.06
ṭu	0.07	ṭhu	0.01	du	0.04	dhu	0.01
tu	0.92	thu	0.00	du	0.99	dhu	0.15
pu	1.58	phu	0.21	bu	0.71	bhu	0.11
ke	6.10	khe	0.13	ge	0.45	ghe	0.03
ṭe	0.24	the	0.11	de	0.76	dhe	0.10
te	2.01	the	1.13	de	2.01	dhe	0.08
pe	0.20	phe	0.08	be	0.52	bhe	0.17
ko	3.80	kho	0.18	go	0.42	gho	0.24
ṭo	0.04	tho	0.03	do	0.14	dho	0.04
to	1.38	tho	0.13	do	0.56	dho	0.03
po	0.06	pho	0.04	bo	0.24	bho	0.08

the number of classes is large and when some classes, by nature, occur rarely. One solution to this problem is to use isolated utterance data for training the models. We consider the SCV data excised from continuous speech for our studies on recognition of frequently occurring SCV classes and the isolated utterance data for the studies on recognition of all the SCV classes.

1.5.5 Varying Durations of SCV Utterances

Durations of the utterances vary for different SCV classes. Durations of segments belonging to a particular SCV class also vary depending on the speaking rate and the context in continuous speech. This suggests that it is necessary to use classifier architectures that can handle varying duration patterns derived from SCV segments. Though the durations of SCVs vary, the durations of some of the regions (burst, aspiration and transition regions) for an SCV do not vary as much. The variations in the overall duration manifest mostly as variations in the durations of the closure and vowel regions. The closure and vowel regions are the initial and final regions of an SCV segment respectively. Because the discriminatory clues in these regions are present mainly in their steady characteristics, it is not necessary to process the complete durations of these regions. This suggests that most of the information necessary for recognition of SCVs can be captured by processing the portion of an SCV segment around the vowel onset point, that contains parts of the closure and vowel regions, and all of the burst, aspiration and transition regions. Therefore it is possible to represent any SCV utterance, irrespective of its overall duration, by a fixed duration pattern retaining most of the relevant information and thus avoid the need for handling varying duration patterns.

1.5.6 Detection of Vowel Onset Points in SCV Segments

In order to obtain fixed duration patterns automatically from varying duration SCV segments, it is necessary to identify the Vowel Onset Point (VOP) in an SCV segment. Once the VOP is identified, a portion of speech signal with a fixed duration around the VOP can be processed to obtain a fixed duration pattern. The method for identification of VOP has to be robust and independent of speaker. In our studies on classification of SCVs, we explore suitable methods for identification of VOPs in SCV segments.

1.5.7 Parametric Representations for SCV Segments

Two important aspects of parametric representation are identification of parameters suitable for different significant regions of an SCV utterance [9] and development of methods for the extraction of the parameters from speech signal. Some main issues related to these aspects are mentioned below:

- Parameters for distinguishing voiced/unvoiced closure regions.

- Parameters for identifying the frequency of the peak in the spectrum of the burst region which is short in duration.

- Parameters to capture the characteristics of an aspiration region when it is present.

- Parameters for representing the transition regions. The transition region may start at the beginning of the vowel region as in unaspirated SCVs or during the aspiration region as in aspirated SCVs. Parameters for the transition region should be able to capture and represent the formant transitions.

- Parameters to represent the formant frequencies and amplitudes of the vowel region, even when the duration of the steady portion of the vowel region is short.

- Methods to absorb the variations in parameters due to speakers and the context, without losing the discriminatory information.

1.5.8 Importance of Transitions in SCV Segments

The characteristics of sounds in SCV segments are affected by the adjacent sounds because of coarticulation effects. This is more true in the case of continuous speech. The clues for recognition of stop consonants are dependent on their immediate context significantly. Formant transition clues for the place of articulation of stop consonants are dependent on the following and preceding vowels [10], [11]. It is necessary to capture this information present in the transition regions of SCV segments for correct classification of SCVs [12]. It is necessary to represent the transition regions by suitable speech parameters that help to capture the information from the time varying speech signal in those regions.

1.5.9 Fuzzy Nature of Clues for Recognition of SCVs

There are many sources of variability in the characteristics of the SCV utterances in continuous speech. The main sources due to the speech production mechanism are the speaker characteristics, the context and the speaking rate. It is difficult to collect and enumerate the variations in different sounds due to each source. The recognition strategies to handle the variability should either isolate the variability by looking for the invariant discriminatory clues or should use the methods to take the variability into account. The variability in the characteristics of the utterances mainly manifests as variations in the parameters extracted from the speech signal. Therefore it is necessary to obtain a representation of the parameters that is less sensitive to the variability. One method for doing this is to transform the parameters into features that describe the parameters. The features can be linguistic descriptions of the parameters. The linguistic descriptions are best represented by fuzzy logic. This suggests the need for fuzzification of speech signal parameters to handle the variability in the utterances. The process of fuzzification has to be done carefully in a way that does not lead to loss of the discriminatory information present in the parameters extracted from the speech signal.

1.5.10 Issues Addressed in the Thesis

We first address the issues in developing neural network models for classification of the SCV segments manually excised from continuous speech for a small set of frequently occurring SCV classes. Then we address the issues in developing models for classification of isolated utterances of all the SCV classes. We develop a method for deriving the fixed duration patterns from varying duration SCV segments and utterances. Parametric representation based on weighted cepstral coefficients is used in deriving the patterns from SCV segments. We explore suitable methods for representing the transition regions in SCVs.

The issues discussed so far are concerned with classification of the SCV segments excised from continuous speech and isolated utterances of SCVs. In the next section, we discuss the main issues in developing models for spotting the SCV segments in continuous speech.

1.6 Issues in Spotting SCV Segments in Continuous Speech

Though there have been many successful efforts in developing approaches for spotting keywords in unrestricted speech, these approaches cannot be extended for spotting subword units. The main reasons are: (1) The vocabulary of keywords is application dependent whereas the number of subword units is language dependent, and (2) Confusability amongst the keywords is not as high as that amongst the subword units. Additionally, the approaches for keyword spotting make the following assumptions: (1) A given continuous speech utterance has only one occurrence of one of the keywords, and (2) A given continuous speech utterance can be modeled as a segment containing one of the keywords that may be preceded and followed by nonvocabulary speech segments. These assumptions that simplify the design of models for spotting keywords are not valid in developing models for spotting subword units. Development of systems for spotting subword units requires classification models capable of discriminating large number of similar classes.

Strategies for spotting subword units in continuous speech have been based on training classification models to classify only the segments of the continuous speech signal belonging to subword units and reject all other segments. The models thus trained to classify or reject are then used to scan the speech signal continuously and hypothesize the presence or absence of the corresponding subword units. This strategy for spotting requires models capable of rejecting all nonvocabulary segments in continuous speech. For spotting SCVs, we consider a strategy in which the vowel onset points (VOPs) are first located in continuous speech and then the scanning by classification models for SCVs is limited to the regions around VOPs. This locate and scan strategy is likely to result in fewer false alarms than the commonly used strategy based only on scanning. The locate and scan strategy spotting requires a robust method for automatic identification of VOPs in continuous speech. In our studies on spotting SCVs, we develop a neural network based method for detection of VOPs in continuous speech and adopt the locate and scan strategy for spotting.

1.7 Organization of the Thesis

A 'roadmap' showing the evolution of ideas reported in this thesis is given in Fig.1.1.

The organization of the thesis is as follows: In Chapter 2, we first review methods for extraction of suitable parametric representations for recognition of SCV segments and then present a review of approaches for automatic recognition of subword units in continuous speech. In Chapter 3, we present our studies on developing models for classification of SCV segments manually excised from continuous speech for a small set of frequently occurring SCV classes. In Chapter 4, modular neural networks based approaches are explored for handling the large number of SCV classes. We consider

Neural Network Models for Recognition of SCV Segments

1. Continuous speech recognition:

- **Signal-to-symbol** transformation: Acoustic-phonetic analysis
- **Symbol-to-text** conversion: Lexical, syntactic and semantic analysis

2. Signal-to-symbol transformation: **Recognition** of **Subword** units

3. Recognition of **subword** units: **CVs** as units, Spotting, Classification

4. Classification of SCVs: Classification models, Parametric representations

5. Classification models: Good discrimination, Handle large number of classes

6. Discriminatory models: Neural network models

7. Neural network models: Fixed duration pattern classifiers

8. Derivation of fixed duration patterns: Vowel onset points in SCVs

9. Classification of SCV segments excised from continuous speech:

- Large number of SCV classes: Similarity amongst several classes
- Varying frequency of occurrence of SCV classes: Limited training data
- Variability in characteristics of SCVs in continuous speech
- Difficulty in capturing all the variations from the limited training data

10. Neural network architectures: OCON and ACON architectures

11. Models for large number of classes: Training data, Modular networks

12. Training data for all SCV classes: Isolated SCV data

13. Derivation of patterns: Effects of segment durations

14. Modular networks: Criteria for grouping classes into subgroups

15. Constraint satisfaction network:

- Neural networks for subsets of classes: Nonlinear feature extractors
- Combination of evidences from multiple sources
- Constraints based on acoustic-phonetic knowledge and experimental data

16. Spotting SCVs: Strategy for spotting, Models for spotting

17. Locate and scan strategy: Vowel onset points in continuous speech

18. Models for spotting: Rejection of non-SCV segments

19. Parametric representations:

- Representation of SCV transitions: Pitch region analysis
- Reduction of variability: Fuzzification

Figure 1.1: Evolution of idem described in the thesis.

different criteria guided by the phonetic description of SCVs for grouping the large number of SCV classes into subgroups. In Chapter 5, we propose a constraint satisfaction model to combine the evidences available from the networks based on different grouping criteria. Studies on spotting SCV segments in continuous speech are presented in Chapter 6. Studies on suitable parametric representations for transitions in SCV segments are described in Chapter 7. Methods for fuzzification of formant trajectories in SCV transitions are also explored in this chapter. Finally, we summarize the contributions of this research work in Chapter 8.

Chapter 2

.

REVIEW OF APPROACHES TO RECOGNITION OF SUBWORD UNITS

In this chapter, we first give a review of approaches based on different classification models for automatic recognition of subword units. Then we present the nature of clues for recognition of Stop Consonant-Vowel (SCV) segments and give a review of methods for extraction of the clues from speech signal. In Section 2.1, we review approaches based on hidden Markov models for classification of subword units such as phonemes and syllables. Approaches based on artificial neural network models are reviewed in Section 2.2. Modular neural networks based approaches for classification are reviewed in Section 2.3. In Section 2.4, we give a review of approaches for spotting subword units in continuous speech. In Section 2.5, we review the methods for extraction of clues for classification of SCV segments. We present a review of methods for detection of vowel onset points in Section 2.6.

2.1 Approaches based on Hidden Markov Models

Statistical pattern recognition approaches for continuous speech recognition are mainly based on hidden Markov models [4] [13]. Hidden Markov Models (HMMs) are used for their inherent ability to incorporate the sequential and stochastic nature of the speech signal. An HMM is a doubly stochastic process with an underlying stochastic process that is hidden, but can only be observed through another set of stochastic processes that generates the sequence of observed symbols [14]. An HMM is characterised by a finite number of states, a finite number of observation symbols per state, a transition (from one state to another) probability distribution, an observation symbol probability distribution and an initial state probability distribution. In discrete HMMs, observations are characterised as discrete symbols and a discrete probability density is used to specify the observation probability distribution. Continuous speech signal representations are converted into a sequence of discrete symbols using vector quantization methods. Continuous HMMs use continuous observation densities to model continuous signal representations directly. The continuous HMMs need larger training data sets because the number of model parameters to be estimated is much larger than that in discrete HMMs.

Approaches based on statistical models for large vocabulary continuous speech recognition have used HMMs for modeling subword units. Models for words are built as concatenations of models for subword units and statistical language models are used for matching at sentence level [14] [15]. Commonly used subword units are context-independent phones, context-dependent phones (diphones and triphones) [16] [17], syllables and acoustic units [18] [19]. In most of these approaches the training of systems is carried out at the level of sentences and the recognition performance of systems is given in terms of word accuracies and sentence accuracies, making them suitable for task specific continuous speech recognition [14]. The recognition performance at word and sentence levels is ultimately limited by the performance at subword unit level. The performance at subword unit level is more important for vocabulary independent, task independent continuous speech recognition. Therefore we will focus on the approaches for recognition at subword unit level.

The recognition performance at subword unit level for a continuous speech recognition system that uses discrete hidden Markov models for phones as subword units is given in [16]. A recognition accuracy of 64.1% for context independent phone units and 73.8% for right context dependent phone units is reported for the system using a bigram phone model as the language model. A recognition accuracy of 58.1% is reported for 8 stop consonant, context independent phone units used in this system.

Recently approaches based on modeling segments within the stop consonants have been used for obtaining an improved recognition performance. HMMs have been developed for stationary microsegments of stop consonants such as silence, voice bar, burst and aspiration and a concatenation of the models for microsegments is used to model a stop consonant [20]. Results of recognition studies indicate an improvement in the performance using this approach over an approach that used a single model for each stop consonant. An HMM representation of quantized articulatory features of consonants and vowels has been used for speaker dependent recognition of 18 isolated stop consonant-vowel and CVC utterances in [21]. In [22], stop consonants are modelled using continuous HMMs as consisting of several well defined microsegments, and a recognition accuracy of 73.6% has been reported for speaker independent recognition of 5 stop consonants in VCV segments excised from continuous speech.

The main limitation of the HMMs in using them for recognition of confusable vocabulary is their poor discriminatory capability [23]. Training HMMs using maximum mutual information (MMI) criterion has been considered for incorporating discriminatory information [24]. But optimization procedures for estimation of HMM model parameters using MMI criterion are complex and often lead to numerical problems in implementation. Approaches based on artificial neural networks have been found to be suitable for discriminatory training. In the next section, we review approaches based on ANNs for recognition of subword units.

2.2 Approaches based on Artificial Neural Network Models

A number of properties that a layered feedforward neural network should have so that it will be useful for speech recognition are listed in [7]. First, the network should have multiple layers and sufficient interconnections between units in each of these layers. This is to ensure that the network will have the ability to learn complex nonlinear decision surfaces. Second, the network should have the ability to represent relationships between events in time. Third, the actual features of abstractions learned by the network should be invariant under translation in time. Fourth, the learning procedure should not require precise temporal alignment of the labels that are to be learned. Fifth, the number of weights in the network should be sufficiently small compared to amount of training data *so* that the network is forced to encode the training data by extracting regularity.

Having listed the properties, the paper describes a time delay neural network (TDNN) architecture for recognition of phonemes {b,d,g} that satisfies these properties. The input to the TDNN consists of 15 frames of speech centered around the handlabeled vowel onset. Each frame consists of normalized mel scale spectral coefficients derived from the speech sampled at 12 kHz, Hamming windowed and analyzed using a 256 point FFT every5 ms. Adjacent coeffcients in time are collapsed to give an overall frame rate of 10 ms. Coefficients are normalized to lie between -1 and +1. The first hidden layer consists of 8 time-delay hidden units. The input to these time-delay units expanded out spatially into a three frame window. In the second hidden layer, each of three time-delay units look at a five frame window of activity levels in the first hidden layer. The-output is obtained by integrating the evidence from each of the three units in the second hidden layer over time and connecting it to its pertinent output unit. The speech data used for training and testing was extracted from the isolated utterances of Japanaese words from three speakers. In performance evaluation, the TDNN achieved an average recognition score of 98.5%.

The reasons for various design choices in the TDNN architecture are discussed in [25]. Each unit in the first of two hidden layers is connected to three successive frames of input (this is called receptive field). This enables the network to capture the relationships between events in time. This is also the reason for calling the architecture as time-delay neural network. To permit the detection of multiple features in each slice of the input, the network has multiple number of hidden units connected to each receptive field. To eliminate the misalignment problem during learning, the network is forced to apply the same set of feature detectors to every slice of the input. Still, there may be incorrect alignment of the utterance during recognition. To solve this problem, a network that contains several copies of each output unit is suggested.

Having described the efforts based on TDNN architectures to develop speech recognition systems, we now summarize other efforts to develop phoneme recognition systems using multilayer neural networks. A connectionist structure for phoneme recognition proposed in [26] has two main parts: (1) A sound subunit classifier using a backpropagation network with two hidden layers to classify speech subunits from frames of speech data and (2) A sequence classifier to classify phonemes from input sequences of subunits by their occurrence and duration. 128 points in the FFT spectrum of speech sampled at 10 kHz and split into nonoverlapping 10 ms frames are used as inputs to the subunit classifier network. The network has 15 output nodes corresponding to the 15 subunits (3 initial stop consonants, 4 final stop consonants, 2 fricatives, 5 vowels and 1 silence). A sequence processor for recognition of phonemes from the subunits is also developed. Overall recognition rate of 87% on the test set of 90 pseudo-words of type C1VC2 or C1D (where C1 is initial consonant, C2 is a final consonant, V is a pure vowel and D is a diphthong) is reported.

Two neural network models, multilayer perceptron and radial basis function network, are applied to a static speech recognition problem in [27]. In the multilayer perceptron network model, the class boundaries are modeled by hyperplanes defined by the hidden nodes. In the radial basis function networks, hidden nodes define hyperellipsoids. Phonetic labeling experiments were conducted on handsegmented vowel tokens of 20 classes (12 monophthongs and 8 diphthongs). The analysis method extracts 20 LPC-derived median cepstral coefficients for each third of the token and 12 coefficients representing a coarse coding of duration. Both the network models are reported to give similar recognition performance of about 70% on test data. The radial basis function network could be trained much faster than the multilayer perceptron network of same complexity.

Nine different parametric representations of speech based on linear predictive parameters are compared in [28]. The input to a multilayer perceptron classifier are parameters extracted for a 20 ms segment excised from the center of steady-state part of a vowel. The MLP classifier is shown to perform best with the cepstral coefficient representation, which gave a recognition score of 91% over 900 utterances of /b/-vowel-/b/ syllables from three speakers.

Two schemes to obtain phonemic transcriptions of spoken utterances are described in [29]. Both schemes utilize the self-organizing Kohonen maps [30] first to vector quantize speech into a sequence of phoneme labels centiseconds apart. In the first scheme, the quasiphoneme sequence is converted into a phoneme string using simple durational transformation rules. In the second scheme, the conversion is carried out by using a multilayered feedforward network. The input vector to the self-organizing maps consists of 15 component approximations of the short time power spectra of the speech signal. Using durational transformation rules, the phonemic accuracy achieved is 83.6%. The feedforward network with one hidden layer is used in the second scheme. The network is trained to filter the transitions between phonemes out of the quasiphoneme sequence. A phoneme recognition rate of 85.5% is achieved using multilayer networks.

Other studies on classification of CVs are mostly on isolated CV utterances as in [31] for English letters, or on isolated CV utterances with the same vowel as in [32] [33] for the Eset (B,C,D,E,G,P,T,V) of the English alphabet.

Many clues for identification of speech sounds and their features are in a linguistic form. They are best represented by using fuzzy logic [34]. Fuzzy logic based approaches have been used for classification of patterns [35]. Recently there have been attempts to combine neural networks and fuzzy logic based approaches for pattern classification [36] [37]. Neural networks that use fuzzy representation of formant data for recognition of vowels have been shown to give a better classification performance in [38]. Fuzzy representation of the similarity of patterns was used in training neural networks for recognition of vowels and it has been shown that the fuzzy neural networks based approaches give better performance than the conventional approaches [39].

2.3 Modular Networks based Approaches

Monolithic neural network architectures are not suitable for developing classifiers for a large number of classes, as in the case of subword unit based continuous speech recognition. Modular networks based approaches have been proposed for recognition of large number of classes. In these approaches modularity is viewed as a manifestation of the principle of divide and conquer, which permits one to solve a complex computational task by dividing it into simpler subtasks and then combining their individual solutions [40]. A neural network is said to be modular if the computation performed by the network can be composed into two or more modules (subsystems or subnetworks) that operate on distinct inputs without communicating with each other. The outputs of the modules are mediated by an integrating unit that is not permitted to feed information back to the modules.

Modularity has been used as a design strategy in developing large phonemic networks for recognition of all consonants [41]. In this approach, several time-delay neural networks have been developed for different subsets of consonants and the outputs of these subsystems are combined to determine the consonant class. Phonemes are grouped into the following subgroups: ({b,d,g}, {p,t,k}, {m,n,syllabic nasal sN}, {s,sh,h,z}, {ch,ts}, {r,w,y} and {a,i,e,o,u}). A discrimination score of above 96% for each of these subgroups is reported. When two of these networks for {b,d,g} and {p,t,k} are combined to build a network for {b,d,g,p,t,k}, the performance decreased to about 60%. This indicated that further training of the combined network is necessary to improve the performance. Different strategies for incremental learning were explored. They are (1) use of class distinctive features, (2) connectionist glue techniques where more hidden units are included in the hidden layer 1 to learn any missing discriminatory features and (3) all-net fine tuning. After combination learning and all-net fine tuning, the consonant network yielded a recognition score of about 95% for the phonemes excised from the utterances of Japanese words. This is compared with an improved version of HMM model which gave a recognition score of 92.7% on the same data.

In another approach, a hierarchical strategy has been proposed for handling the large number of phoneme classes [42]. In this strategy, the subgroup of a given input pattern is first decided by a network and the subsystem for that group is permitted to classify the given input pattern into one of the classes in its corresponding subgroup. Recently, modular approaches have been considered for recognition at subword unit level in hybrid HMM/ANN based approaches for continuous speech recognition [43].

2.4 Approaches to Spotting Subword Units

Though there have been many successful efforts in developing approaches for spotting keywords in unrestricted speech, these approaches cannot be extended for spotting subword units. The main reasons are: (1) The number of keywords in the vocabulary is application dependent and is much smaller than the number of subword units which is language dependent and (2) Confusability amongst the keywords is not as high as that amongst the subword units. Additionally, the approaches for keyword spotting make assumptions that a given continuous speech utterance has only one occurrence of one of the keywords, and that the given continuous speech utterance can be modeled as a segment containing one of the keywords that may be preceded and followed by non-vocabulary speech segments. These assumptions that simplify the design of models for spotting keywords are not valid in developing models for spotting subword units.

Hidden Markov models have been extensively used in keyword spotting systems because of their ability to model keywords of varying durations and also because they can be used to build models that satisfy the above assumptions [44] [45] [46]. Development of subword unit spotting systems requires classification models capable of discriminating large number of similar classes. Discriminant techniques for training HMMs used for word spotting have been shown to improve the spotting performance [47]. A learning algorithm based on discriminative learning theory, namely, Minimum Classification Error formalization/Generalized Probabilistic Descent method(MCE/GPD), has been proposed for minimizing errors in spotting five Japanese consonants [48].

Artificial neural networks are shown to have a better discriminatory capability than HMMs. Approaches based on ANNs for spotting words have used self-organizing map and feed-forward networks [49], recurrent networks [50], neural tree networks

[51] and multiple Restricted Coloumb Energy networks [52]. Time delay neural networks (TDNN) have been considered for spotting phonemes and a small set of CV syllables [42] in word utterances. In order to develop a vocabulary independent continuous speech recognition system by spotting subword units in continuous speech, it is necessary to evolve suitable strategies for minimizing errors in spotting subword units which are large in number and which form a highly confusable set of classes. We address the issues in developing approaches for spotting subword units of Stop Consonant-Vowel (SCV) classes.

2.5 Methods for Processing SCV Segments

Stop consonants are considered to be the most difficult consonants to recognize for the following reasons [53]: (1) The speech production mechanism of a stop consonant is dynamic, involving a closure and release period, (2) The complex nature of this production mechanism results in many diverse acoustic cues, and (3) The acoustic events during the production of the sound can be omitted or severely distorted. In this section, we briefly describe the speech production mechanism of stop consonants, identify the important clues for recognition of stop consonants and present the nature of these clues.

2.5.1 Speech Production Mechanism for Stop Consonants

Production of stop consonants are characterized by the following successive significant events[54]: (1) Closure that can be voiced or silent, (2) Transient corresponding to the response of the vocal tract to the pressure release, (3) Frication that is characterized by noise produced at the consonantal constriction, (4) Aspiration characterized by an 'h-like' noise, and (5) Transition corresponding to the initial part of a following voiced sound to the extent that it is influenced by coarticulation with the stop. For stop consonants in English and many other European languages, the speech segment corresponding to the transient, frication and aspiration events is treated as a single segment called the 'burst'. In Indian languages, the presence or absence of the aspiration event is one of the discriminating characteristics of different stop consonant sounds. Therefore the speech segment corresponding to the transient and frication events is considered as 'burst' for stop consonants in Indian languages. Different significant events in production of an SCV utterance /**kha**/ in Hindi are shown in Fig.2.1. The figure shows the plots of speech signal waveform, formant frequencies, and the regions of different significant events in the utterance.

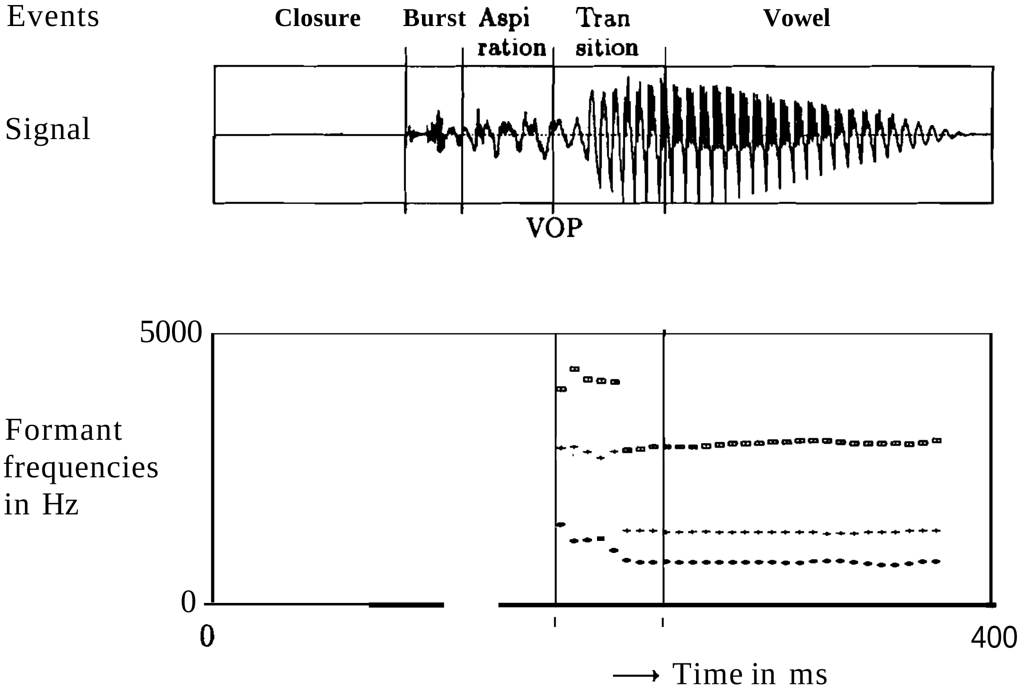


Figure 2.1: Different significant events in the production of the SCV utterance /**kha**/. The figure shows the plots of signal waveform and formant frequencies, and indicates the boundaries of regions corresponding to different significant events in production of the SCV utterance. The vowel onset point (VOP) is also indicated.

Stop consonants in Hindi are characterized by their manner of articulation and stop consonants in English and many other European languages, the speech segment corresponding to the transient, frication and aspiration events is treated as a single segment called the 'burst'. In Indian languages, the presence or absence of the aspiration event is one of the discriminating characteristics of different stop consonant sounds. Therefore the speech segment corresponding to the transient and frication events is considered as 'burst' for stop consonants in Indian languages. Different significant events in production of an SCV utterance **/kha/** in Hindi are shown in **Fig.2.1.** The figure shows the plots of speech signal waveform, formant frequencies, and the regions of different significant events in the utterance.

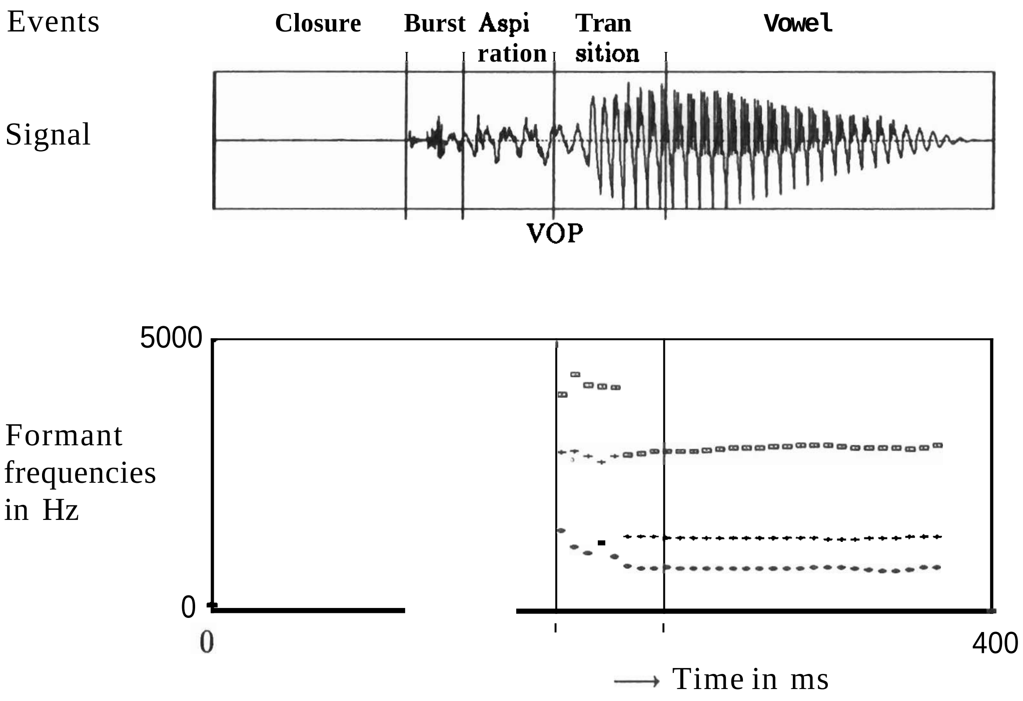


Figure 2.1: Different significant events in the production of the SCV utterance **/kha/.** The figure shows the plots of signal waveform and formant frequencies, and indicates the boundaries of regions corresponding to different significant events in production of the SCV utterance. The vowel onset point (VOP) is also indicated.

Stop consonants in Hindi are characterized by their manner of articulation and place of articulation [55]. There are four different manners of articulation and four different places of articulation. Typical speech signal waveforms for different stop consonants of Hindi are shown in Fig.2.2. The plots for the **16** stop consonants are arranged in four columns and four rows. The four columns correspond to the four manners of articulation and the four rows correspond to the four places of articulation. The points at which the vocal tract closure occurs for different places of articulation [56] are also shown in Fig.2.2. Speech signal waveform plots show the closure, burst, and aspiration regions (wherever they are present), and the initial portions of the vowel for utterances of SCVs with the vowel /a/. In the remainder of this section, we present the important clues and nature of these clues for identification of manner and place of articulation of stop consonants.

2.5.2 Clues for Manner of Articulation of Stop Consonants

Manner of articulation of stop consonants is described by the unvoiced or voiced nature of the closure event and the presence or absence of the aspiration event, leading to four different manners of articulation. They are (1) Unvoiced-Unaspirated, (2) Unvoiced-Aspirated, (3) Voiced-Unaspirated and (4) Voiced- Aspirated. The main clues for recognition of the manner of articulation are present in the segments corresponding to the closure and aspiration events. The closure segment is characterized by a low energy region and the aspiration segment by a formant structure similar to that of the following vowel, but with no periodicity [55]. The acoustic features corresponding to these clues are [57] [55]: (1) Voicing during closure, (2) Voice onset time, (3) Nature of first formant transition, (4) Spectral flatness and (5) Ratio of the high frequency energy to the low frequency energy. Because of incomplete articulation of sounds in continuous speech, voicing during closure may have a weak presence and aspiration may also have a weak presence or it may completely be absent. Recogni-

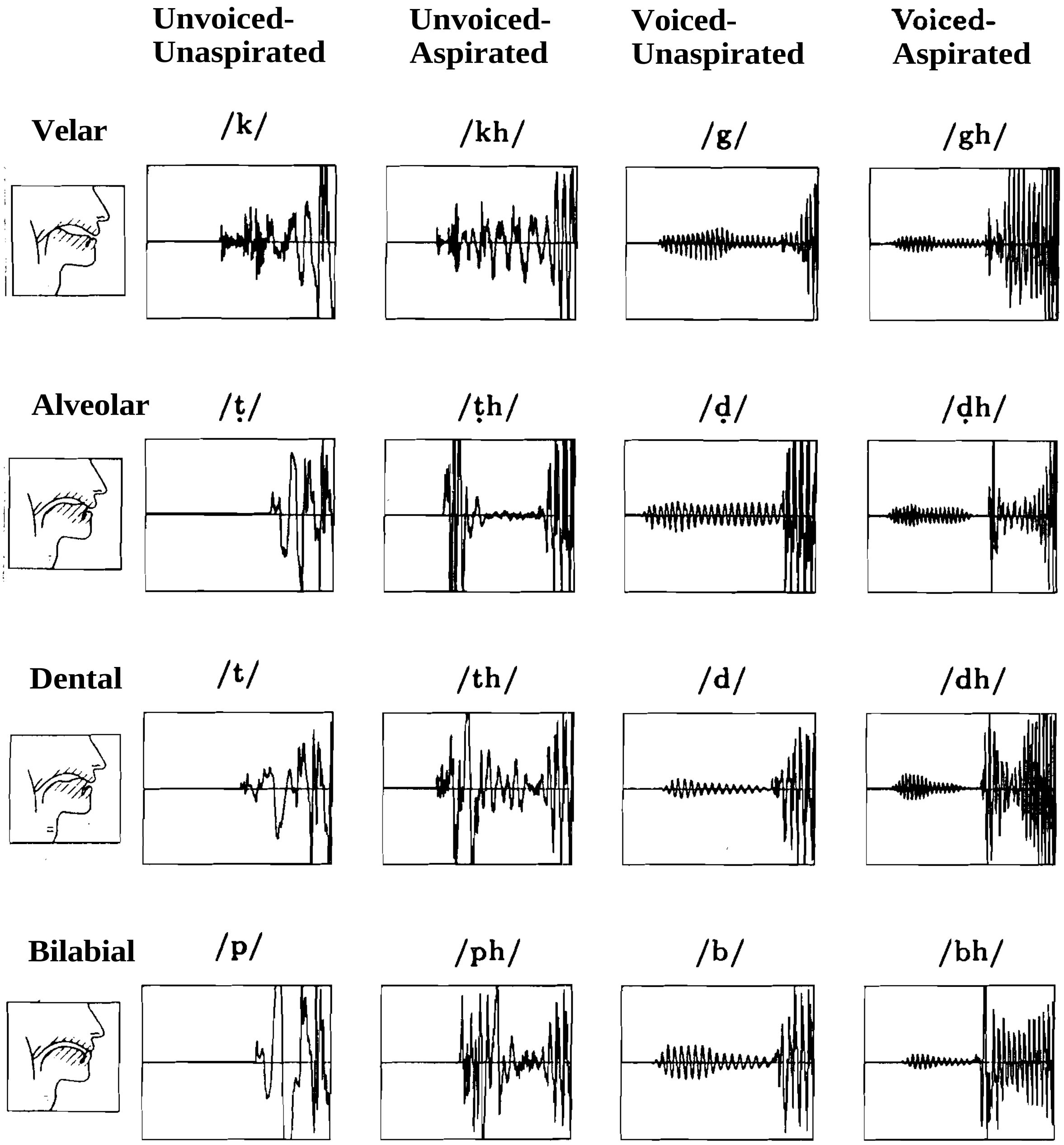


Figure 2.2: Different stop consonants in Hindi. The figure shows the plots of speech signal waveforms for 16 stop consonants in Hindi. The plots are arranged in 4 columns and 4 rows. The 4 columns correspond to the 4 manners of articulation of Hindi stop consonants. The 4 rows correspond to the 4 places of articulation. The point in the vocal tract at which the closure occurs for each place of articulation is also shown. Speech signal waveform plots show the closure, burst, and aspiration regions (wherever present), and the initial portions of the vowel for utterances of SCVs with the vowel /a/.

tion errors (such as /dha/ being classified as /da/) due to imprecise articulation of the manner of stop consonants can be handled only by using lexical knowledge in the symbol-to-text conversion stage of a speech recognition system.

2.5.3 Clues for Place of Articulation of Stop Consonants

Place of articulation of stop consonants is described by the portion of the vocal tract at which the consonantal constriction occurs during production of stop consonant sounds. There are four places of articulation for stop consonants in Indian languages, namely, (1) Velar, (2) Alveolar, (3) Dental and (4) Bilabial. The main clues for recognition of the place of articulation are present in the segments corresponding to the burst and transition events [58]. The burst segment is characterized by a short duration (5-10 ms) and the transition segment is characterized by a dynamic spectrum. The acoustic features corresponding to the clues for place of articulation are [57] [59]: (1) Distribution of energy in the burst spectrum, (2) Dynamics in the burst spectrum, and (3) formant transitions between the release of stop consonants and steady region of the following vowels. Development of suitable signal processing methods for extraction of the features from speech segments of short durations and time varying spectral characteristics is important for recognition of the place of articulation of stop consonants. In continuous speech, the burst may have a weak presence or may even be absent. Even when it is present, it is difficult to identify the burst segment and extract the features only from the burst segment using the existing signal processing methods. Formant transition features are certainly present in any SCV segment. Therefore they are more reliable and important features than the burst segment features [60] [61]. The formant transition features for a stop consonant are dependent on the adjacent vowels [10] [62]. Therefore it is necessary to take the vowel context into account for extraction of formant transition features. We next review methods for modeling and processing the transition segments.

Modeling Transitions in SCVs

There are mainly two methods for modeling the transitional behaviour in consonantal environments. The first method is based on the locus theory [63] where it is assumed that for each consonant there is a single target spectrum or locus with the property that in VC and CV transitions the vowel spectra tend to converge towards the target for the consonant [64]. The identification of loci associated with place of articulation from acoustic analyses of real speech has been elusive. In a recent attempt, the locus concept has been generalized to that of locus equations that describe linear relationships between the formants at the voicing onset of CV syllables and those at the midvowel nuclei [65]. It has also been shown that locus equations have the property of relational invariance, i.e., they are invariant with respect to the consonantal place and are relational with respect to the vowel context. Context dependent hidden Markov models structured by locus equations have been developed for modeling and classification of transitions in a CVC environment [66] [67].

The second method for modeling transitional behaviour in consonantal environments is based on acoustic measurements aimed at quantifying formant transition patterns in relation to the vowel formant values [11]. In this method, the transition regions in CV and VC segments are analyzed to extract features related to formant frequencies and changes in formant frequencies. The focus in this method is on developing suitable signal processing techniques for extraction of the formant features from transition regions characterized by dynamic spectral characteristics. Processing techniques based on a time-varying model [68] for speech signal have been shown to give a better classification performance than the standard processing techniques for recognition of unvoiced stops in VC environments [69]. In our studies, we consider pitch region analysis based processing techniques for extraction of features from transition regions SCV segments [70].

2.6 Methods for Detection of Vowel Onset Points

In this subsection, we present a review of methods for detection of Vowel Onset Points (VOPs) in continuous speech. The objective of using a method for detection of VOPs in recognition of SCVs is to focus on processing and analysis of speech segments around VOPs. These segments contain most of the relevant information necessary for recognition of SCVs. In classification of SCVs using neural network models, segments with a fixed duration around VOPs are processed to obtain fixed duration patterns which are given as input to the neural network classifiers. For spotting SCVs, scanning of speech signal can be limited to only the segments around VOPs. This helps in automatically eliminating many portions of the continuous speech signal that do not contain SCVs and in reducing the number of false alarms during spotting.

One of the methods for detection of VOPs in continuous speech is based on segmentation of continuous speech into vowel and nonvowel like regions [71]. In this method speech signal parameters such as energy, ratio of the high frequency energy (energy in the range of 3800-4600Hz) to the low frequency energy (energy in the range of 0-800Hz), ratio of the volumes of back and total cavities of vocal tract and ratio of the volumes of front and back cavities of vocal tract were used as features to discriminate vowel and nonvowel like regions. The performance of this method on 10 sentences containing about 90 VOPs was shown to give an error percentage of about 3% on male speaker data and about 7% on female speaker data.

A method for detection of VOPs in continuous speech by identifying the points at which there is a rapid increase in the vowel strength was proposed in [72]. In this method the vowel strength is computed using the difference in energy of each of the peaks in the amplitude spectrum and the energy of a dip associated with the peak. Speech segments with duration of pitch periods are analysed to obtain the amplitude spectrum and computing the vowel strengths. This method was shown to give a correct detection performance of 91% on speech data containing about 375 VOPs with a precision of 20 msec or better.

Another method for detection of VOPs is based on first classifying the speech signal into voiced/unvoiced/silence regions [73] [74] and then labelling the voiced regions as vowel and nonvowel regions. A method for voiced/unvoiced/silence classification by automatic labelling of instants of significant excitation in speech signal has been proposed in [75]. In this method absolute and relative parameters of speech signal energy and linear prediction residual energy of frames with a duration of about 3 ms and on both the sides of significant points of excitation [76] were used to train a feedforward network for classification of speech signal into voiced/unvoiced/silence regions.

In our studies, we consider two methods for detection of VOPs. One method is concerned with detection of VOPs in SCV segments excised manually from continuous speech and in isolated utterances of SCVs. This method is based on detection of the point where there is a rapid increase in the energy of the speech signal. This method is used in our studies on classification of SCVs. The second method is concerned with detection of VOPs in continuous speech. This method is used in our studies on spotting SCVs in continuous speech.

2.7 Summary

Approaches based on neural networks for classification of subword units were developed for the isolated utterances of units or for the segments manually excised from isolated words and continuous speech. Patterns derived from handsegmented portions of a fixed duration were used for training and testing the classification models. In our studies, we develop a method for- automatically deriving the fixed duration patterns from varying duration segments. This method is based on detection of vowel onset points and processing fixed duration segments around them.

Phonemes have been used as subword units in many approaches for continuous speech recognition. The number of phonemes is small but it is difficult to recognize them because of the coarticulation effects on their characteristics in continuous speech. Linguistic constraints at word and sentence level have been used to correct the errors in recognition of phonemes. Syllables have not been used mainly because they are large in number. We have chosen CVs as basic units. We develop approaches based on modular neural networks for handling the large number of units. We consider different criteria for grouping the unit classes into subgroups and training a separate network for each subgroups. We develop a constraint satisfaction model that uses the acoustic-phonetic knowledge of the classes to combine the evidences from modular networks for different groupings. This approach for classification of subword units is vocabulary independent.

Spotting subword units is important for vocabulary independent continuous speech recognition. Approaches used for spotting keywords are not suitable for spotting subword units. Many approaches for spotting have been based on scanning the speech signal of a sentence continuously. We address the issues in developing an approach for spotting based on the detection of vowel onset points in continuous speech and scanning only the segments around them. The focus of this thesis work is on vocabulary independent recognition of the stop consonant-vowel (SCV) segments.

Chapter 3

STUDIES ON CLASSIFICATION OF SCV SEGMENTS

3.1 Objectives of the Studies

The objectives of the studies presented in this chapter are: (1) to develop an approach based on neural network models for classification of the SCV segments excised from continuous speech, (2) to compare the performance of different models and architectures for classification of the SCV segments, and (3) to analyze the performance of the models in order to identify the main sources of errors in classification. In the next section, we present an approach for classification of the SCV segments. In Section 3.3 and 3.4, we describe the classification models and architectures used in our studies. The classification studies and the results are presented in Section 3.5. In Section 3.6, we give an analysis of the results of the studies.

3.2 An Approach for Classification of SCV Segments

The SCV segments in continuous speech have varying durations. The durations are approximately in the range of 75 to 350 ms. It is observed that SCVs with short vowels /i/ and /u/ and occurring at the end of phrases and sentences have short durations. The segments belonging to the SCV classes with long vowel /a:/ in the beginning of phrases and sentences have been observed to be of long durations. Neural network classifiers considered in our studies are capable of handling only fixed duration patterns. Therefore it is necessary to derive fixed duration patterns from SCV segments of varying durations. The fixed duration patterns should have all the important information necessary for classification. In this section, we develop an approach for deriving fixed duration patterns..

An SCV segment consists of all or a subset of the following significant speech production events: Closure, Burst, Aspiration, Transition and Vowel. Any SCV segment will have the regions corresponding to the closure, transition and vowel events. The burst is supposed to be present for all the stop consonants. In **SCVs** occurring in continuous speech, the bursts may be totally absent or may have a weak presence. The aspiration is supposed to be present in the aspirated stop consonant sounds. It is observed that in continuous speech the aspirated stop consonants are some times pronounced as unaspirated sounds or they are characterised with weak aspiration. The important and reliable clues for classification are present in the closure, transition and vowel regions. It is important to take these aspects into account in development of models for classification of the SCV segments excised from continuous speech.

The signal waveform and the formant frequencies for two segments belonging to the classes **/ka/** and **/ka:/** are shown in Fig.3.1. It is important to note that though the difference in the total durations of the two segments is high, the durations of the events such **as** burst and transition are not much different. The difference in the total durations mainly manifests **as** differences in the durations of the closure and vowel regions.

The closure and the vowel regions are the initial and final regions of an SCV segment respectively. Because the discriminatory features in these regions are present mainly in their steady characteristics, it is not necessary to process the complete durations of these regions. The information necessary for classification can be captured

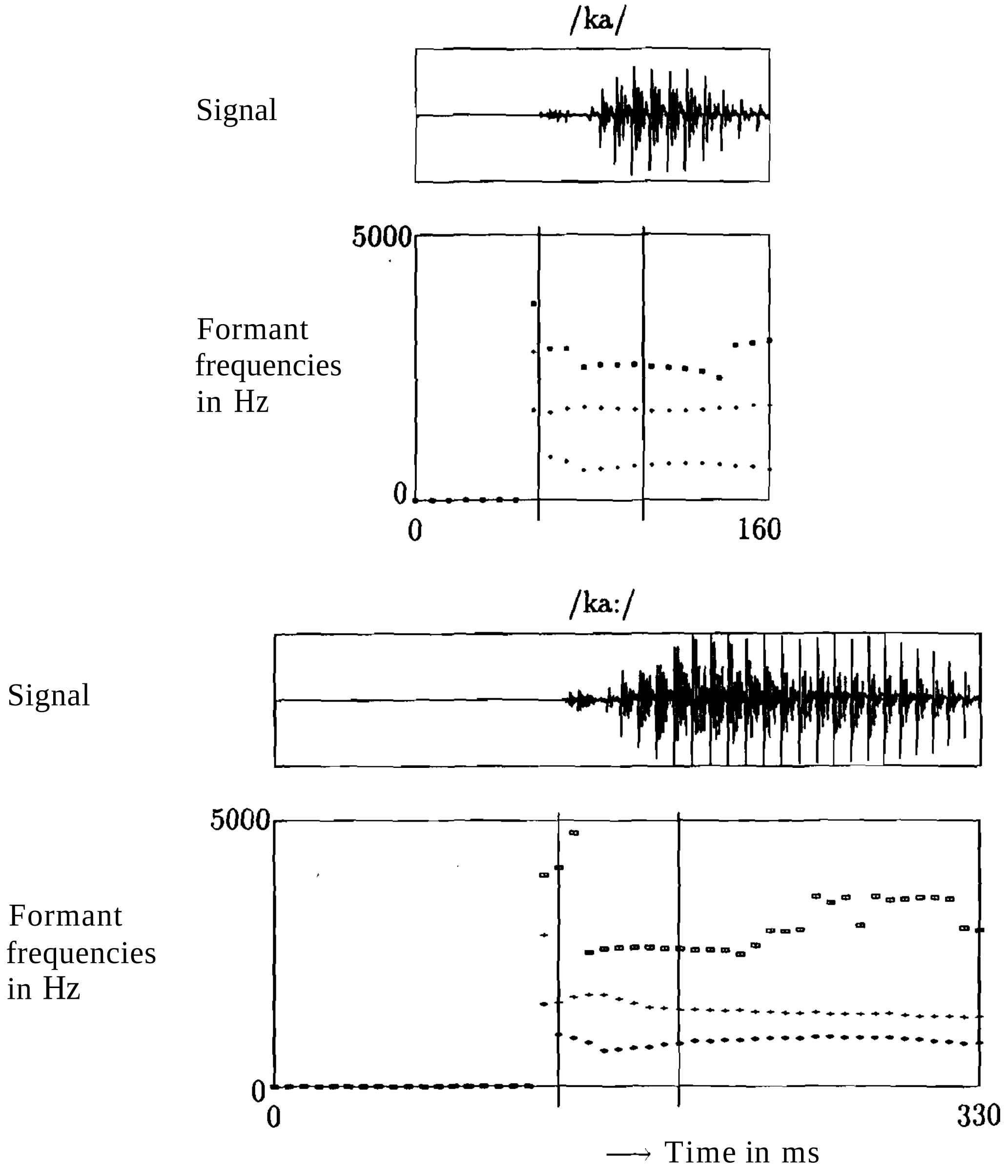


Figure 3.1: Differences in the durations of significant events for SCVs with short and long vowels. The figure shows the plots of signal and formant frequencies for the SCV segments of /ka/ with short vowel and /ka:/ with long vowel. The transition regions are marked in the plots of formant frequencies. It can be seen that though there is a difference of 170 ms in the overall durations of /ka/ and /ka:/, this difference mainly manifests as differences in the durations of the closure and vowel regions. The durations of the burst and transition regions are not affected significantly.

by processing the portion of an SCV segment containing parts of the closure and vowel regions, and all of the burst, aspiration and transition regions. The closure, burst and aspiration regions are present before the Vowel Onset Point (VOP) in any SCV segment. The transition and vowel regions are present after the VOP. Since it is possible to automatically identify the VOP in an SCV segment, a fixed duration signal around the VOP that contains most of the necessary information can be processed to derive a fixed duration pattern.

In our studies, we consider a 100 ms long signal starting at 20 ms before the VOP and ending at 80 ms after the VOP. The signal before the VOP would include the closure, burst and aspiration events that may be present in an SCV segment, and the signal after the VOP would include the transition and vowel regions. These durations are arrived at after observing that the duration of transition regions is in the range of 30-40 ms and that the characteristics of the vowel in an SCV are to be captured by processing at least 30 to 40 ms of the steady portion of the vowel region. We now present a method for the detection of VOPs in the SCV segments.

3.2.1 Detection of Vowel Onset Points in SCV Segments

It is important to identify the vowel onset points in the SCV segments with a good accuracy because they form the anchor points around which the signal is processed to derive the patterns. An error in detection of the VOP leads to deriving a pattern that does not include the necessary information for classification. In this section, we consider a method based on the derivative of the signal energy to identify the VOPs in the SCV segments.

It is observed that the energy of the signal increases rapidly at the VOP in the SCV segments. This is because the energy of the signal immediately after the VOP is much higher than that in the closure or burst regions that immediately precede the

VOP. Though in most of the SCV segments the increase in energy is the highest at VOPs, it is observed that in some cases the energy continues to increase even after the VOP and the point at which the maximum increase in energy occurs does not coincide with the VOP. The proposed method computes the maximum of the energy derivative in the SCV segment and then identifies the first point from the beginning of the SCV segment at which the derivative is above a threshold. The threshold is a fraction of the maximum energy derivative value. The signal waveform, the energy derivative and the VOP identified are shown in Fig.3.2 for a segment belonging to the class /kha/. It is observed that in some of the aspirated SCVs, the beginning of the aspiration region is identified as VOP because there is a significant increase in the energy when the closure region ends and the aspiration region begins. This is illustrated in Fig.3.3 a segment belonging to the class /dha/. It is necessary to evolve a better method for detection of VOPs in the aspirated SCVs.

Once the VOP in an SCV segment is identified, a 100 ms long signal around the VOP is processed to obtain 20 frames with 12 weighted cepstral coefficients in each frame. The weighted cepstral coefficients are derived from an 8th order linear prediction analysis [77], using a frame size of 20 ms and a shift of 5 ms. The algorithm used for extraction of weighted cepstral coefficients is given in Appendix A. If an SCV segment does not have 20 ms signal before the VOP or 80 ms after the VOP, then the first and the last frames of the segment are duplicated to derive a 20 frame pattern. Duplicating the frames in the steady regions of the SCV segments does not affect their classification.

The method proposed in this section has been used to obtain the training and test patterns of the SCV data used in our studies. In the next section, we describe the models used for classification of the SCV segments.

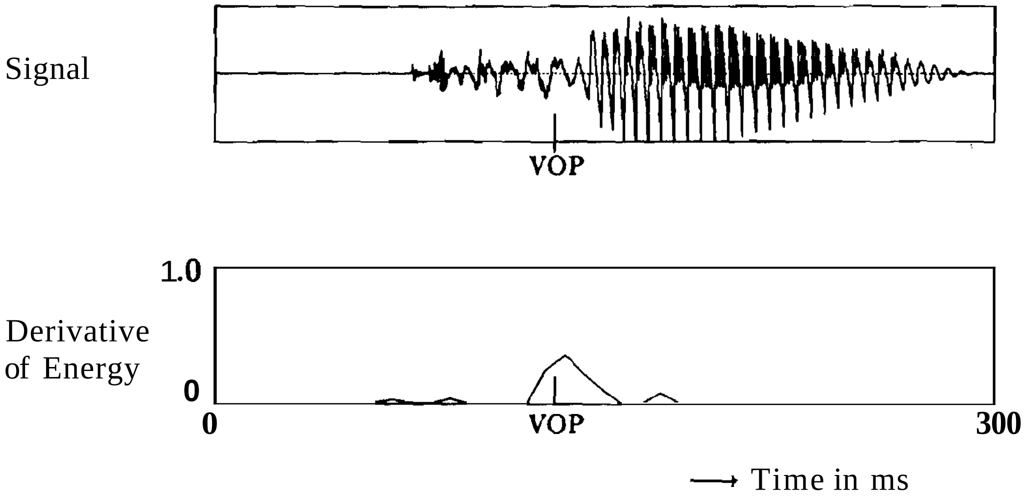


Figure 3.2: Detection of vowel onset point (VOP) in an SCV segment. The figure shows the plots of signal waveform and the derivative of energy. The largest peak in the derivative of energy occurs at the VOP.

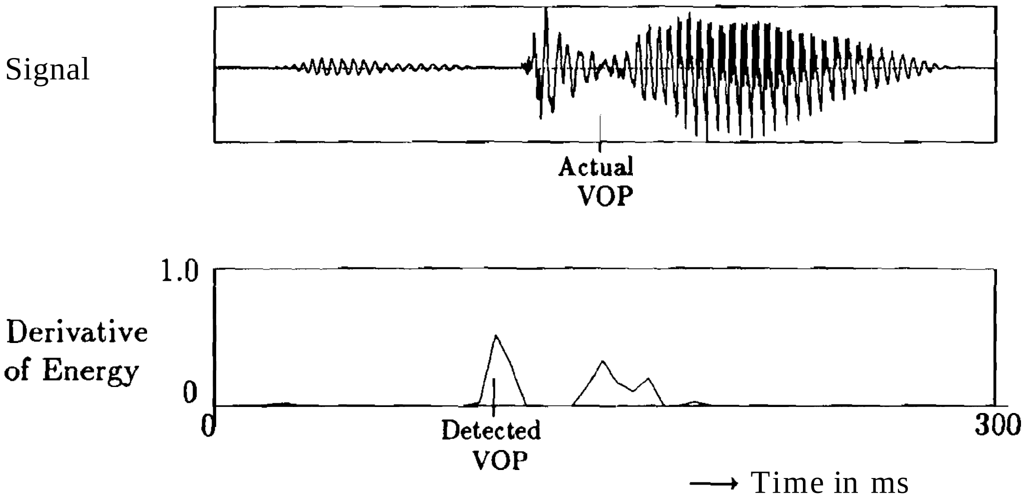


Figure 3.3: Misdetection of vowel onset point (VOP) in an SCV segment. The figure shows the plots of signal waveform and the derivative of energy. The largest peak in the derivative of energy occurs at the beginning of aspiration region instead of occurring at the VOP. Therefore the detected VOP is different from the actual VOP shown.

3.3 Neural Network Models for Classification of SCVs

3.3.1 Multilayer Perceptron

Approaches based on artificial neural networks have been used for speech recognition [6]. The main advantages of neural networks are a powerful discrimination based learning procedure and relatively mild assumptions about statistical distributions [40]. Multilayer perceptrons are the commonly used neural networks. Multilayer perceptrons with two hidden layers are capable of forming complex decision surfaces using hyperplane bounded decision regions spread across multiple layers. It has been shown that multilayer perceptrons give a better generalization performance compared to statistical models when the underlying distributions of classes are not known [78]. The structure of the multilayer perceptron model used in our studies is shown in Fig.3.4. The input layer consists of **240** nodes to input a **20** frame pattern with **12** weighted cepstral coefficients per frame. A column of nodes in the input layer represents a frame. Each unit in the first hidden layer is connected to each of the **240** nodes in the input layer. Similarly each node in the second hidden layer is connected to all the nodes in the first hidden layer, and each of the nodes in the output layer is connected to all the nodes in the second hidden layer. Standard error backpropagation algorithm [40] is used for training the multilayer perceptron networks.

3.3.2 Time Delay Neural Network

One of the main limitations of the multilayer perceptron model is its inability to provide invariance for translation in time. Time Delay Neural Network (TDNN) [25] model can be used to overcome this limitation. Typical structure of a TDNN used for classification of SCVs is shown in Fig.3.5.

The input to the TDNN is a **20** frame pattern derived from an SCV segment. Each unit in the hidden layer is connected to a certain number of frames of input

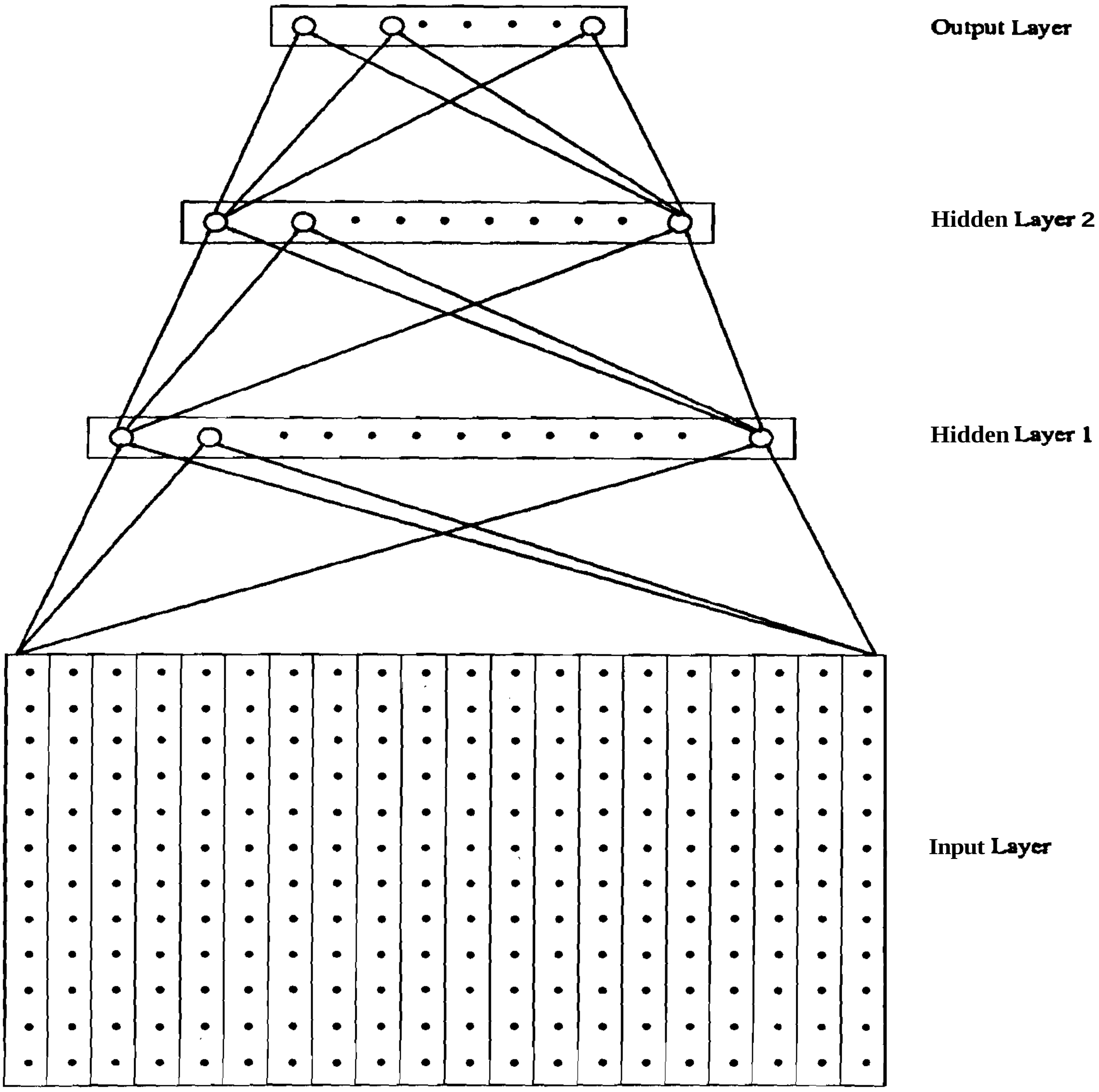


Figure 3.4: Structure of the multilayer perceptron network used for classification of SCVs. The input layer consists of 240 nodes to input a 20 frame pattern with 12 weighted cepstral coefficients per frame. A column of nodes in the input layer represents a frame. Each unit in the first hidden layer is connected to each of the 240 nodes in the input layer. Similarly each node in the second hidden layer is connected to all the nodes in the first hidden layer, and each of the nodes in the output layer is connected to all the nodes in the second hidden layer.

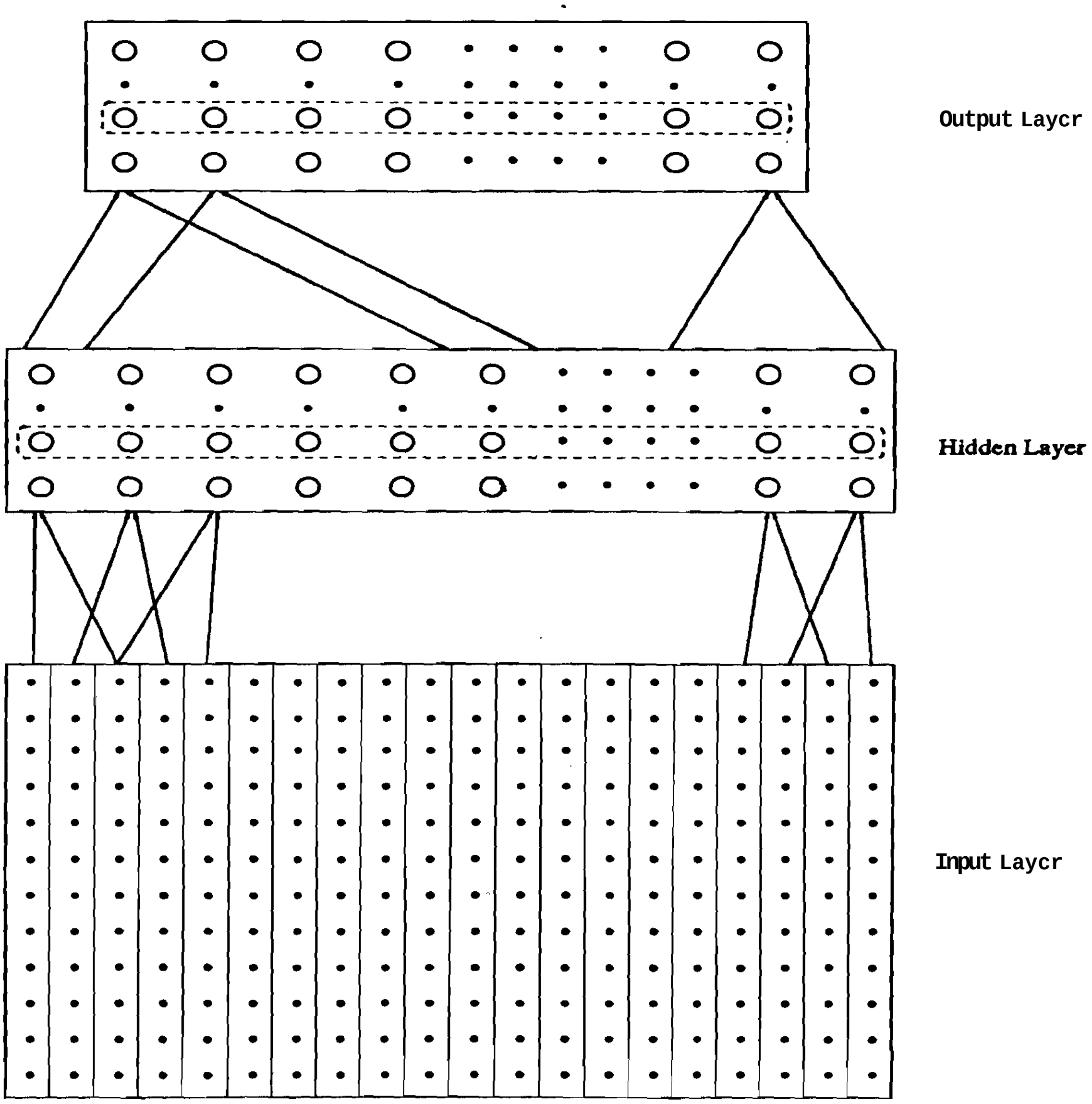


Figure 3.5: Structure of the time delay neural network used for classification of SCVs. The input layer consists of 240 nodes to input a 20 frame pattern with 12 weighted cepstral coefficients per frame. A column of nodes in the input layer represents a frame. Each unit in the hidden layer is connected to 3 consecutive frames of input forming its receptive field. The hidden layer consists of replicas for each hidden node. The number of replicas is same as the number of receptive fields of size 3 in the input layer. The replicas for a hidden node are shown in a row. The number of rows in the hidden layer corresponds to the number of hidden nodes. The replicative structure is also used for the output layer and each output node is connected to a receptive field of size 5 in the hidden layer.

(called receptive field). The number of receptive fields is dependent on the size of the receptive field and the overlap between adjacent receptive fields. For example, the total number of receptive fields is 18 for an input layer consisting of 20 frames, with a receptive field size of 3 frames and an overlap of 2 frames. The hidden layer consists of replicas for each hidden node. The number of replicas for a hidden node is same as the number of receptive fields in the input layer. The number of columns in the hidden layer corresponds to the number of receptive fields in the input layer. The number of rows in the hidden layer corresponds to the number of hidden nodes. Several hidden nodes are used to permit detection of multiple features in each receptive field. Replicas of the nodes are used to detect the same features in different receptive fields.

To eliminate the misalignment problem during training, the replicas of a node use the same set of weights. Still, there may be incorrect alignment of during recognition. To solve this problem, the output nodes are replicated and each replica is connected to a different slice of the hidden layer. The learning algorithm used for training the multilayer perceptron is modified to train the TDNN. The learning algorithm for TDNN is derived in Appendix B.

Because of the similarity amongst SCV classes, it is necessary to incorporate the discriminatory information in the classification models. In the next section, we consider different neural network architectures for incorporating the discriminatory information.

3.4 Neural Network Architectures for Classification of SCVs

We consider two neural network architectures, namely, (1) One-Class-One-Network (OCON) and (2) All-Class-One-Network (ACON), in developing classifiers for the SCV segments. In the OCON architecture a separate network is trained for: each class. The network of a class is trained with patterns belonging to that class which are used as positive examples, and also with patterns belonging to the classes close to it which are used as negative examples. The network is trained to give a high output value for positive examples and a low value for negative examples. The aim of training a network for a class is to form a decision boundary around the region of that class in the pattern space. During classification a test pattern is input to the networks of all the classes and the outputs of the networks are processed to determine its class. The main advantage of the OCON architecture is that the size of the networks is not large. Another advantage is that it is possible to use a suitable preprocessing method for each class. The main disadvantage is that it is difficult to train the network of a class to give a low output value for patterns belonging to many other classes and hence the discriminatory capability of the network can be poor. The structure of the classifier based on the OCON architecture is shown in Fig.3.6.

in the ACON architecture a single network is trained for all the classes. The number of output nodes in the network is same as the number of classes. The structure of the classifier based on the ACON architecture is shown in Fig.3.7. The training data consists of a number of patterns belonging to each class. The network is trained to give a high value for the output node belonging to the class of a training pattern and a low value for all other output nodes. The aim of training is to form decision surfaces among the regions of all the classes in the pattern space. The shapes of the decision surfaces become more complex as the number of the classes increases. It may be difficult to train a single network for large number of classes. If a network can be trained for a given set of classes, the discriminatory capability of the network is expected to be better than that of the OCON architecture. The disadvantage is that it is not possible to use different preprocessing methods for different classes.

In our studies, we consider the OCON and ACON architecture based neural network classifiers using MLP and TDNN models. In the next section, we describe the

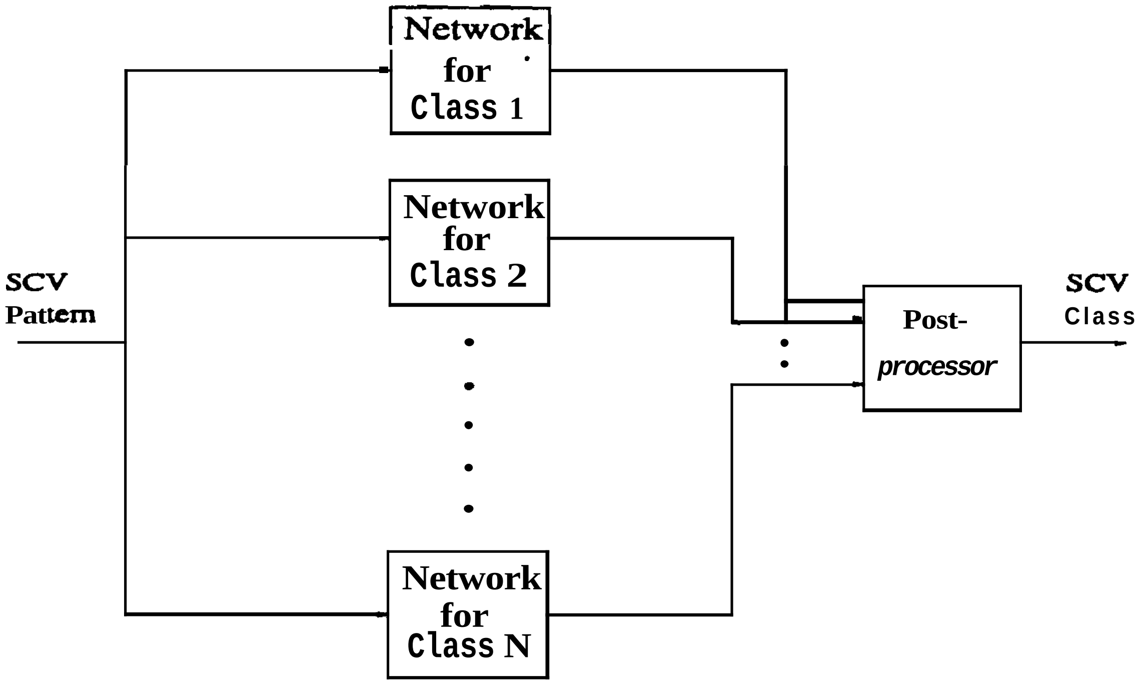


Figure 3.6: One-Class-One-Network(OCON) architecture for classification of SCVs. A separate network is trained for each of the classes under consideration. The number of networks is same as the number of classes. For classification of an SCV pattern, it is input to each network and the outputs of all the networks are combined by a postprocessor that implements the classification criterion to determine its class.

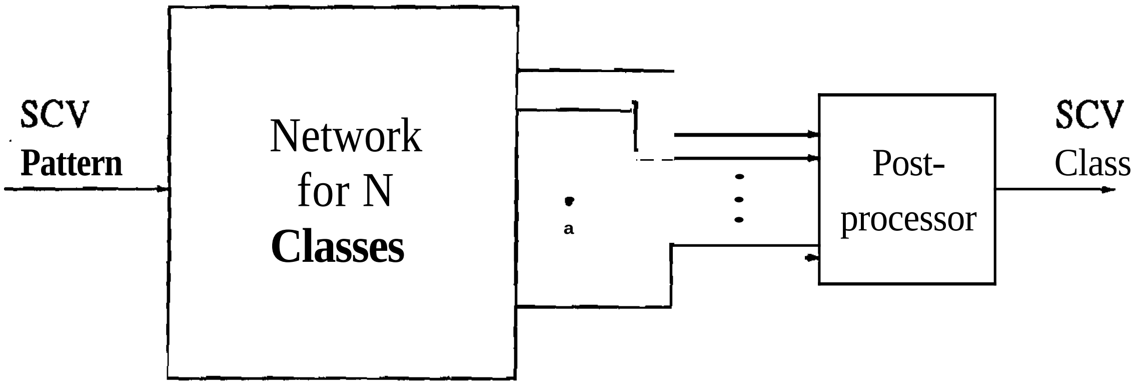


Figure 3.7: All-Class-One-Network (ACON) architecture for classification of SCVs. A single network is trained for all classes under consideration. The outputs of the network for a given SCV pattern are combined by a postprocessor that implements the classification criterion to determine the class of the input pattern.

studies on classification of the SCV segments excised from continuous speech.

3.5 Classification Studies and Results

We have conducted studies on classification of continuous speech segments belonging to a subset of SCV classes in Hindi. In this section we first present the classes of SCVs considered in our studies, the details of the speech data collected and the implementation details of the classifiers. Then we present the studies and their results.

3.5.1 Classes of SCVs for Studies

Speech signal data belonging to each of the 80 SCV classes was collected by manually excising the SCV segments from continuous speech of 50 sentences for each of the 8 speakers (5 male and 3 female) considered in our studies. The total number of SCVs present in this data is about 2500. About 55% of the total number of SCVs belong to the subset of the most frequently occurring six SCV classes (/ka/, /ki/, /ke/, /ta/, /dha/ and /pa/) and about 75% of the total data belong to the subset of ten SCV classes (/ka/, /ki/, /ke/, /ko/, /ta/, /ti/, /to/, /da/, /dha/ and /pa/). The number of segments belonging to the remaining 70 SCV classes is only about 25% of the total data. We have considered the set of six SCV classes in our first study and the set of ten SCV classes in our second study. Though the number of classes considered in the studies is much smaller than the total number of classes, classification of these frequently occurring SCVs is still a challenging problem because these classes are highly confusable.

3.5.2 Implementation Details of Classification Models

We have conducted classification studies using multilayer perceptron (MLP) and time delay neural network (TDNN) models. Models with different number of nodes in the hidden layers have been considered in our studies. The performance is given for the models with the optimal number of hidden nodes. The performance of the models has not improved even if the number of the hidden nodes is increased beyond the optimal number.

For the OCON architecture based classifiers, the MLP modes has 20 nodes in the first hidden layer and 15 nodes in the second hidden layer. The TDNN model has 10 nodes in the hidden layer. For the ACON architecture based classifiers, the performance is given for two MLP models and two TDNN models. The first MLP model, denoted as **MLP1,** has 125 nodes in the first hidden layer and 60 nodes in the second hidden layer. The second MLP model, **MLP2,** uses 20 nodes in the first hidden layer and 15 nodes in the second hidden layer. The first TDNN model, **TDNN1,** has 20 nodes in the hidden layer (same as in the first hidden layer of **MLP2).** The second TDNN model, **TDNN2,** has only 10 nodes in the hidden layer. Training of models was carried out until the total sum of squares error [40] is small and does not change from one epoch to another. It was observed that the TDNN models required much longer training periods compared to the MLP models. This is mainly because of the replicative structure of the TDNN models.

In order to compare the performance of neural network classifiers with that of hidden Markov models, we have considered discrete hidden Markov models (DHMMs). A 5-state, left-bright, discrete HMM was used to model an SCV segment. The structure of this model is shown in Fig.3.8. It is expected that different states would represent different significant events in the production of an SCV segment. Skipping of states is allowed to model the absence of specific events in some SCV utterances. Standard Baum-Welch reestimation method was used for training the DHMMs and the forward procedure was used for recognition [14]. The algorithms for the training and recognition methods are given in Appendix C. Vector quantization of the weighted

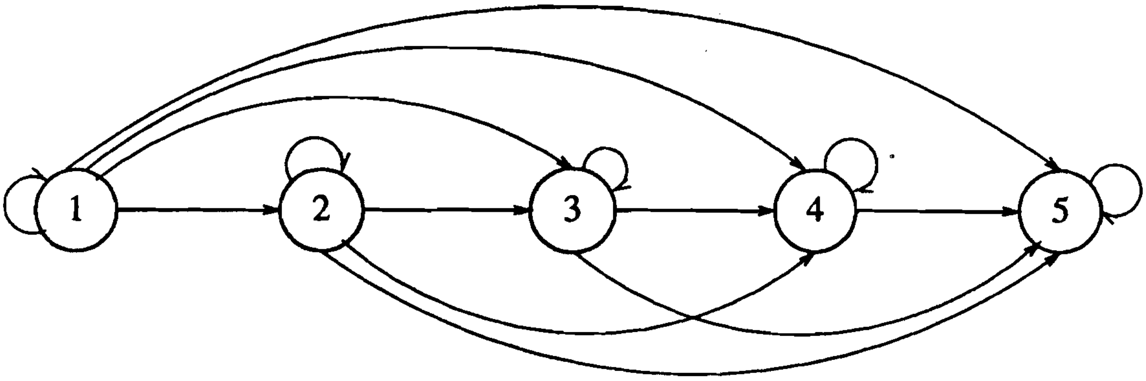


Figure 3.8: Discrete hidden **Markov** model used for classification of **SCVs.** A 5-state, left-to-right model is trained for each SCV class. A circle in the figure represents a state and an arc represents a state transition. The input to the model is a sequence of **codebook** indices corresponding to the weighted cepstral coefficient vectors extracted from speech signal of an SCV segment.

cepstral coefficient vectors was performed using the binary split algorithm [14] to build a **codebook of** size 256.

3.5.3 Classification Studies

Here we present our studies on comparison of the performance of different models and architectures for classification of the SCV segments. The list of the studies carried out are given in Table-3.1.

In our studies, the performance of the classifiers is given for the following two cases of classification criterion: (1) Correct class is the class with the largest output value and (2) Correct class is amongst the classes with the largest and the second largest output values. The second case is considered because it is observed that the class with the second largest output is the correct class in many instances of errors in classification. A significant increase in the performance for the second case suggests that two alternative class symbols can be given for each SCV segment during classification. The correct symbol can be chosen using the lexical and syntactic knowledge

Table **3.1:** List of studies on classification of the SCV segments excised from continuous speech.

1. Comparison of the performance of classifiers based on OCON and ACON architectures, and using MLP, TDNN and DHMM models for six frequently occurring SCV classes.
2. Confusability among the six SCV classes.
3. Comparison of the performance of classifiers based on different models and architectures for ten frequently occurring SCV classes.
4. Confusability among the ten SCV classes.
5. Effects of the inclusion of additional classes on the performance of the OCON and ACON architecture based classifiers.
6. Analysis of the performance of classifiers to identify the main sources of errors in classification.

of the language.

Our first study considers six frequently occurring SCV classes: (/ka/, /ki/, /ke/, /ta/, /dha/ and /pa/). The training set for a network in the OCON classifiers consists of 10 patterns belonging to its class from each of the five speakers. These patterns are used as positive examples. The training set also includes 50 patterns belonging to the other classes which are used as negative examples. The training set for the network in the ACON classifiers consists of 10 patterns per class for each of the five speakers. Thus, a total number of 300 patterns were used for training. The remaining data of the five speakers (582 patterns) was used as the multispeaker test set. The complete data for the six classes from three new speakers (490 patterns) was used as the speaker independent test set. The performance for Case-1, the first case of the classification criterion, is given in Table-3.2(a) and the performance for the Case-2, the second case of the classification criterion, is given in Table-3.2(b) . The performance is given as the percentage of the total number of patterns in a data set that are correctly classified.

Table **3.2;** Comparison of the performance of classifiers based on different models and architectures for six frequently occurring SCV classes. Performance is given for the two cases of classification criterion: (a) Case-1 that an input pattern is correctly classified if the class of the input pattern is the class with the largest value amongst the outputs of a classifier, and (b) Case2 that an input pattern is correctly classified if the class of the input pattern is amongst' the classes with the largest and the second largest output values. It can be seen that the ACON classifiers give a better performance than the OCON classifiers. Classifiers based on the MLP model give a better performance than the TDNN or HMM models. The increase in performance for Case2 over Case-1 indicates that for many patterns that are incorrectly classified, the class with the second largest output value is the correct class.

(a) Performance of classifiers for Case-1 of the classification criterion.

Data Set	OCON		ACON				DHMM
	MLP	TDNN	MLP1	MLP2	TDNN1	TDNN2	
Training	**86.0**	**72.7**	**98.0**	**90.3**	**94.0**	**93.7**	**99.0**
Multispeaker	**75.1**	**71.3**	**86.6**	**84.2**	**81.2**	**80.0**	**71.3**
Speaker-independent	**63.3**	**47.5**	**68.2**	**65.5**	**61.0**	**63.5**	**59.2**

(b)Performance of classifiers for Case2 of the classification criterion.

Data Set	OCON		ACON				DHMM
	MLP	TDNN	MLP1	MLP2	TDNN1	TDNN2	
Training	**98.7**	**94.0**	**98.7**	**96.3**	**98.0**	**99.0**	**100.0**
Multispeaker	**94.0**	**92.6**	**95.0**	**94.2**	**94.5**	**93.8**	**90.9**
Speaker-independent	**82.9**	**75.7**	**85.5**	**79.2**	**84.7**	**83.7**	**83.5**

It is observed that there is a significant increase (by about **10** to **25%)** in the performance on the test data sets for Case-2 over Case-1 of the classification criterion. This indicates that when there are errors in classification, the class with the second largest value is likely to be the correct class. The better performance of the ACON classifiers over the the OCON classifiers indicates the better discriminatory capability of the ACON classifiers. It is also observed that **MLP** models give a better performance than the TDNN or DHMM models.

The best performance is given by the ACON classifier using the **MLPl** model. The performance of this classifier for Case-l'is used to obtain confusion matrices that indicate the confusability amongst the six classes. The confusion matrix based on the performance for the multispeaker test set is given in Table-3.3(a). The confusion matrix based on the performance for the speaker independent test set is given in Table-3.3(b).

Table **3.3:** Confusion matrices for six frequently occurring SCV classes. The confusion matrix based on the performance of the ACON classifier using the **MLPl** model for the multispeaker test set is given in (a) and for the speaker independent test set is given in (b). The entries in a row give the total number of patterns for a class and the number of these patterns that are classified as patterns belonging to each of the six classes. It can be seen that the confusion for a class is mainly with the classes that are phonetically close to it.

(a)Confusion matrix based on the performance for the multisoeaker test data set.

Class	Total	ka	ke	ki	ta	dha	pa
ka	162	151	0	0	2	2	7
ke	53	2	34	14	1	1	1
ki	131	0	2	129	0	0	0
ta	72	3	0	0	58	4	7
dha	92	0	0	4	7	79	2
pa	72	3	0	0	11	5	53

(b) Confusion matrix based on the performance for the speaker independent test data set.

Class	Total	ka	ke	ki	ta	dha	pa
ka	118	92	0	0	8	6	12
ke	59	6	21	13	3	15	1
ki	102	4	9	87	0	2	0
ta	67	10	0	0	31	14	12
dha	8	1 2	0	0	6	66	7
pa	63	3	0	0	9	14	37

The diagonal entries in the matrices give the number of patterns that are correctly classified. The non-diagonal entries in a row give the number of patterns belonging to a class that are misclassified as patterns belonging to each of the other classes. The confusion matrices indicate that in case of errors, the patterns belonging to a class are assigned to the classes which are phonetically close to that class.

Our second study considers four additional SCV classes (/ko/, /ti/, /to/ and /da/). The training set for this study contains additionally five patterns per class of these four classes for each of the five speakers. The remaining data of the ten classes for the five speakers (713 patterns) was used as the multispeaker test set. The, complete data for the ten classes from three new speakers (617 patterns) was used as the speaker independent test set. The performance of different classifiers is given in Table-3.4(a) for Case-1 and in Table-3.4(b) for Case_2.

Table 3.4: Comparison of the performance of classifiers based on different models and architectures for ten frequently occurring SCV classes. It can be seen that the ACON classifiers give a better performance than the OCON classifiers. The models with larger number of hidden nodes (MLP1 and TDNN2) give a better performance. There is a significant increase in the performance for Case2 over Case-1.

(a) Performance of classifiers for Case-1 of the classification criterion.

	OCON		ACON				
Data Set	MLP	TDNN	MLP1	MLP2	TDNN1	TDNN2	DHMM
Training	82.8	69.8	98.5	100.0	91.5	84.0	98.8
Multispeaker	67.0	58.6	73.5	71.2	72.4	67.6	68.3
Speaker-independent	52.4	41.8	61.9	52.8	56.1	50.4	52.1

(b) Performance of classifiers for Case2 of the classification criterion.

	OCON		ACON				
Data Set	MLP	TDNN	MLP1	MLP2	TDNN1	TDNN2	DHMM
Training	97.3	87.3	99.3	100.0	97.5	94.3	100.0
Multispeaker	87.9	81.1	89.6	84.2	89.3	84.4	85.0
Speaker-independent	75.9	65.5	77.6	71.6	71.5	70.5	**70.3**

The observations made from the results in the first study hold good for the second study also. There is a decrease in the performance of a classifier for the ten classes compared to that of the corresponding classifier for the six classes. This is mainly because some of the classes that are included for the second study are close to the classes considered in the first study. This leads to more confusability amongst the classes. An observation made from the performance of the ACON classifiers is that when the complexities (in terms of the number of hidden nodes) of the MLP and TDNN models are approximately the same (MLP2 and TDNN1), the TDNN model gives a marginally better performance compared to the MLP model. The best performance is given by the ACON classifier using the MLP1 model. The confusion matrix derived from the performance of this classifier for the multispeaker and the speaker independent test sets is given in Table-3.5(a) and Table-3.5(b) respectively.

In this section we have presented the studies on classification of SCV segments excised from continuous speech using different models and architectures. In the next section we give an analysis of the performance of different classifiers.

3.6 Analysis of the Performance of Classifiers

The results of the studies presented in the previous section indicate that the best classification performance is 73.5% on the multispeaker test data and is 61.9% on the speaker independent test data for ten frequently occurring SCV classes. The performance of ACON classifiers has been evaluated for models with different number of hidden nodes. It is observed that increasing the number of hidden nodes up to a limit results in an improvement of the performance. The MLP models with more than 125 nodes in the first hidden layer did not show any improvement. The performance of the models with larger number of hidden nodes is limited by the available training data set. As the number of hidden nodes is increased, the TDNN model requires longer training periods and more importantly larger training sets. For example, a

Table 3.5: Confusion matrices for ten frequently occurring SCV classes. The confusion matrix based on the performance of the ACON classifier using the **MLP1** model for the multispeaker test set is given in (a) and for the speaker independent test set is given in (b). The entries in a row give the total number of patterns for a class and the number of these patterns that are classified as patterns belonging to each of the six classes. It can be seen that the confusion for a class is mainly with the classes that are phonetically close to it.

(a)Confusion matrix based on the **performance** for the multispeaker test data set.

Class	Total	ka	ke	ki	ko	ta	ti	to	da	dha	pa
ka	162	134	1	3	2	11	0	1	0	4	6
ke	53	0	37	14	0	0	1	0	1	0	0
ki	131	0	17	96	0	0	18	0	0	0	0
ko	19	2	0	1	13	0	1	1	0	1	0
ta	72	4	0	0	0	53	0	4	0	3	8
ti	33	1	0	6	0	1	22	0	0	2	1
to	29	0	0	0	2	4	0	21	0	2	0
da	50	0	1	0	0	0	0	1	35	13	0
dha	92	0	0	1	0	4	2	0	18	65	2
pa	72	3	0	0	1	11	0	1	5	3	48

(a)Confusion matrix based on the performance for the speaker independent test data set.

Class	Total	ka ke ki ko ta ti to da dha pa
ka	118	1 0 5 0 0 5 0 0 1 0 0 7
ke	59	1 1 2 1 1 3 1 2 1 0 2 8 0
ki	102	4 8 7 2 0 0 1 4 0 0 4 0
ko	21	1 0 0 1 5 0 0 3 0 2 0
ta	67	22 0 0 5 2 2 0 1 1 1 15
ti	38	1 1 6 0 0 2 7 0 0 3 0
to	29	7 0 0 2 2 0 1 4 0 3 1
da	39	7 0 0 2 0 0 0 7 2 2 1
dha	81	3 0 0 6 3 0 0 1 1 5 8 0
.	63	5 0 0 4 4 0 0 4 5 41

TDNN model with even 50 nodes in its hidden layer could not be trained with the available training data set. This is because of the multiplicativeincrease in the number of replicated nodes as the number of hidden nodes is increased.

An analysis of the performance of the OCON and ACON classifiers for the six and ten SCV classes is carried out to identify the effects of including additional classes. We focus on the performance of the **classifiers** using **MLP** models that gave the best performance. In Table-3.6 below we compare the performance of the classifiers on the data belonging to the six SCV classes considered in our first study.

Table **3.6:** Comparison of the performance of the OCON and ACON classifiers for the six and ten frequently occurring SCV classes on the test data sets of the six classes. The inclusion of the additional four classes (**/ko/,/ti/,/to/** and **/da/**) in the ten class classifier did not **affect** the performance of the OCON classifiers for the classes other than **/dha/**. A significant difference in the performance of ACON classifiers can be noted for all the six classes.

SCV Class	Multispeaker Test Data				Speaker Independent Test Data			
	Six Class OCON	Ten Class OCON	Six Class. ACON	Ten Class ACON	Six Class OCON	Ten Class OCON	Six Class ACON	Ten Class ACON
ka	**87.0**	**81.5**	**93.2**	**82.7**	**66.4**	**64.7**	**77.3**	**88.2**
ki	**57.3**	**54.2**	**98.5**	**73.3**	**71.6**	**71.6**	**85.3**	**70.6**
ke	**75.5**	**75.5**	**64.2**	**69.8**	**40.7**	**40.7**	**35.6**	**35.6**
ta	**84.7**	**84.7**	**80.6**	**73.6**	**20.9**	**20.9**	**46.3**	**32.8**
dha	**73.9**	**65.2**	**85.9**	**70.7**	**84.0**	**50.6**	**81.5**	**71.6**
pa	**72.2**	**69.4**	**73.6**	**65.3**	**82.5**	**82.5**	**58.7**	**65.1**
Average	**75.1**	**71.1**	**86.6**	**74.7**	**63.3**	**57.3**	**68.2**	**65.1**

It is observed that there is a significant difference in the performance of the six class and the ten class OCON classifiers only for the class **/dha/**. It is interesting to note that there is a difference in the performance of the ACON classifiers for all the six classes. The performance of the ten class ACON classifier is significantly less than that of the six class ACON classifier for the classes **/ki/** as well as **/dha/**. The different behaviours of the OCON and ACON classifiers when additional classes are included can be explained as follows.

Each of the networks in the **OCON** classifiers is trained to form a decision boundary around the region of its class in the pattern space. The networks in the **ACON** architecture are trained to form decision surfaces amongst the regions of all the classes. Inclusion of additional classes may not alter the decision boundaries formed for the **OCON** networks whereas it may significantly alter the decision surfaces of the **ACON** classifiers. This is indicated in the confusion matrices of the six class and the ten class **ACON** classifiers given in Tables 3.3 and 3.5. These tables show that the decrease in the performance of the ten class **ACON** network for the class /ki/ is due to the presence of the class /ti/ which differs from /ki/ only in the place of articulation of the stop consonant. The difference in the performance for each of the six classes is because of the readjusted decision surfaces when additional classes are included.

The behaviour of the **OCON** and **ACON** classifiers is illustrated in Fig.3.9 for an arbitrary 2-dimensional pattern space. Typical boundaries expected to be formed for two classes by networks in the **OCON** classifier are shown in Fig.3.9(a). The boundaries for these classes do not change when the number of classes is increased to four as shown in Fig.3.9(b). This is because a network is trained for each class separately. The boundaries for the same two classes expected to be formed by the network in the **ACON** classifier are shown in Fig.3.9(c). It can be noted that the boundaries for these classes change when the number of classes is increased to four as shown in Fig.3.9(d).

It is observed that both the **OCON** and **ACON** classifiers show a decrease in the performance for the class /dha/. This is found to be mainly due to the presence of the class /da/ in set of ten **SCV** classes. Most of the errors in classification of /dha/ segments that are classified as /da/ have been found to be due to the absence of aspiration in their production. The difference in the phonetic descriptions of the

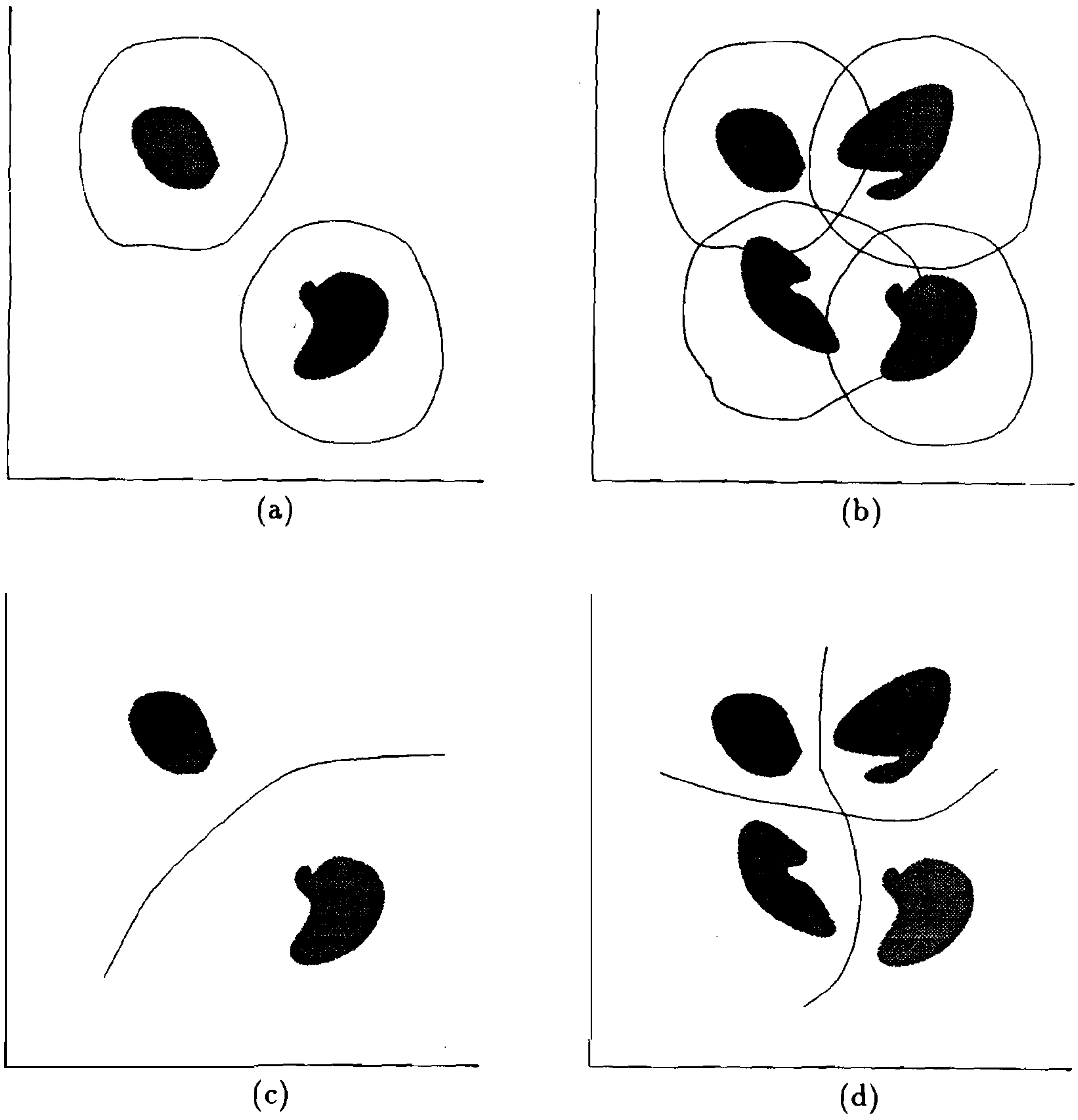


Figure 3.9: Illustration of decision boundaries formed by the OCON and ACON classifiers. An arbitrary 2-dimensional pattern space is used to explain the effect of inclusion of additional classes. **Typicai boundaries** expected to be formed for two classes by networks in the OCON classifier are shown in (a). The boundaries for these classes do not change when the number of classes is increased to four as shown in (b). This is because a network is trained for each class separately. The boundaries for the same two classes expected to be formed by the network in the ACON classifier are shown in (c). It can be noted that the boundaries for these classes change when the number of classes is increased to four as shown in (d).

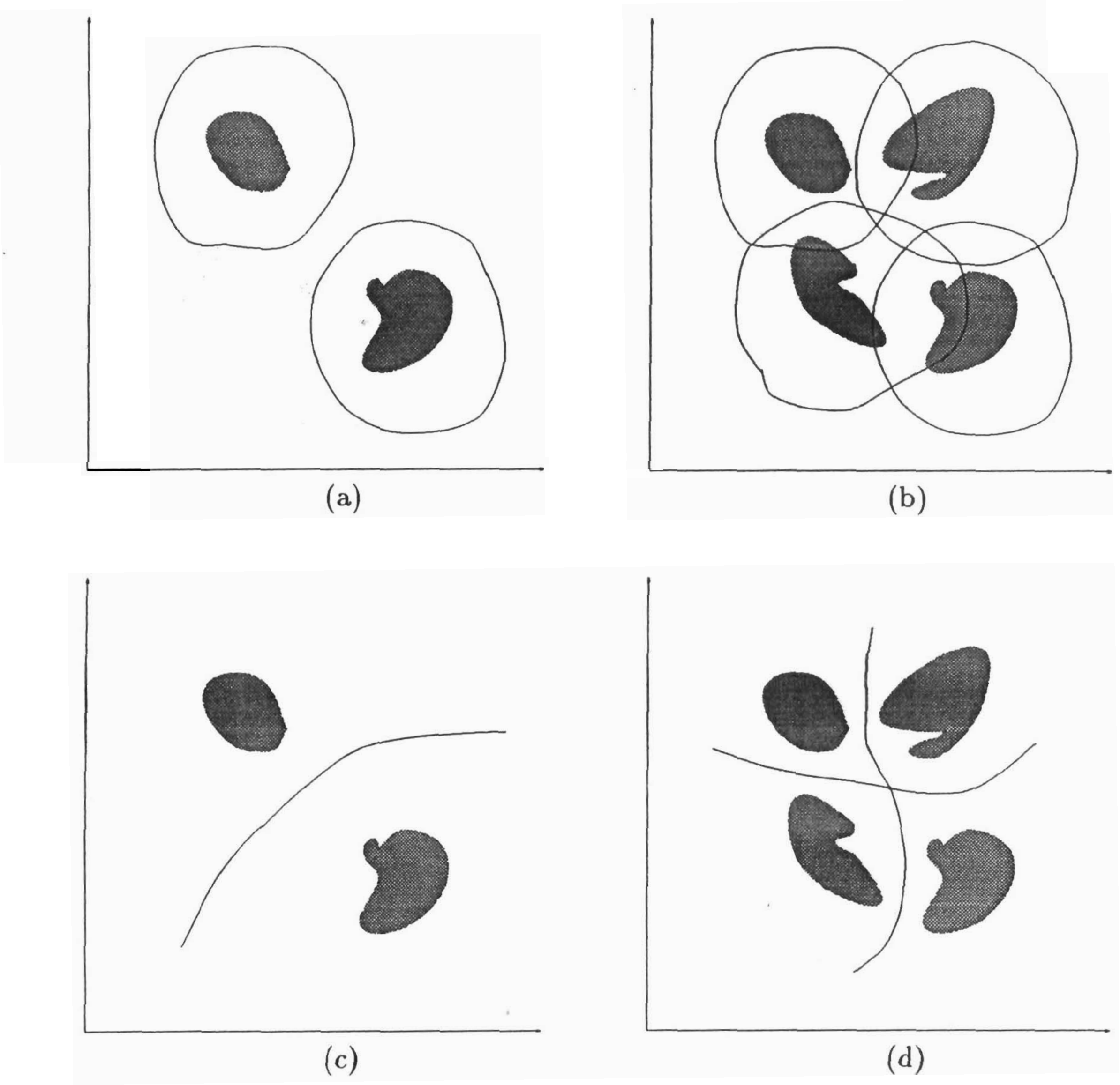


Figure 3.9: Illustration of decision boundaries formed by the OCON and ACON classifiers. An arbitrary 2-dimensional pattern space is used to explain the effect of inclusion of additional classes. Typical boundaries expected to be formed for two classes by networks in the OCON classifier are shown in (a). The boundaries for these classes do not change when the number of classes is increased to four as shown in (b). This is because a network is trained for each class separately. The boundaries for the same two classes expected to be formed by the network in the ACON classifier are shown in (c). It can be noted that the boundaries for these classes change when the number of classes is increased to four as shown in (d).

Chapter 4

MODULAR NEURAL NETWORKS FOR LARGE NUMBER OF CLASSES

4.1 Introduction

In the previous chapter, we have developed a method for classification of SCV segments excised from continuous speech. The performance of different classification models was evaluated for a small set of frequently occurring SCV classes. Neural network architectures considered for a small set of classes have limitations in extending them for large number of classes. In this chapter, we consider approaches based on modular neural network architectures for all the SCV classes which are large in number. The classifiers developed for all the SCV classes can be used to spot any SCV segment in continuous speech.

The main reason for limiting the studies presented in the previous chapter to a small set of classes is the difficulty in collecting adequate training set data from continuous speech for infrequently occurring classes. We consider the isolated utterances of SCVs so that the required training set data can be collected for all the classes. Another important reason for considering the isolated utterance data is that the variability in the characteristics of the isolated utterances is expected to be less compared to the variability in the characteristics of the SCV segments in continuous speech. This is because the isolated utterances are well articulated and they are not affected by the factors such as the context and speaking rate which significantly affect the characteristics of the segments in continuous speech. The variability due to these classes /da/ and /dha/ is that /da/ is unaspirated whereas /dha/ is aspirated. But in continuous speech, many aspirated sounds are produced as unaspirated sounds. Therefore the difference in the phonetic description may not have manifested in the signal. It is difficult to train the networks to form decision boundaries among the overlapping regions. The classification errors due to imprecise articulation of sounds can be corrected using the lexical and syntactic knowledge of the language only. The above analysis indicates that the performance of the OCON classifiers is less affected than the performance of the ACON classifiers when additional classes are included. However, the ACON classifiers have been shown to have a better discriminatory capability for a given set of classes compared to the OCON classifiers.

An analysis of the performance of the classifiers for ten SCV classes segments was carried out to determine the distribution of errors due to misclassification of each of the following: (1) place of articulation (POA) of consonants only, (2) manner of articulation (MOA) of consonants only, (3) vowel only, (4) POA and MOA, (5) POA and vowel, (6) MOA and vowel, and (7) POA, MOA and vowel. The distribution of errors for the multispeaker test data set is given in Table-3.7(a) and the distribution for the speaker independent test data set is given in Table-3.7(b). The total number of errors for a model is shown in the parentheses and the distribution of errors is given as the percentage of the total number of errors. It is observed that a significant percentage (about 25% to 40%) of the errors is only due to misclassification of the place of articulation, irrespective of the classification model used. This indicates the need for strategies to classify the POA more accurately in order to obtain an improved performance. In Chapter 7, we present the studies that address the issues in classification of the place of articulation.

Table 3.7: Distribution of errors in the performance of classifiers for ten frequently **occurring** SCV classes due to misclassification of one or more of the following three features of an SCV: (1) Manner of articulation (MOA) of the consonant, (2) Place of articulation (POA) of the consonant, and (3) Vowel. The distribution for the multispeaker test data is given in (a) and the distribution for the speaker independent test data is given in (b). The entries in the parantheses indicate the total number of errors. A significant percentage of errors is due to misclassification of POA, irrespective of the classification model used.

(a)Distribution of errors in the performance of classifiers on the multispeaker test data set of 713 patterns.

	OCON		ACON				
Source of Errors	MLP (235)	TDNN (295)	MLP1 (189)	MLP2 (205)	TDNN1 (197)	TDNN2 (231)	DHMM (226)
POA only	26.0	43.8	37.1	31.2	24.4	38.1	38.1
MOA only	21.3	16.6	20.1	23.4	32.0	24.2	19.0
Vowel only	29.4	19.3	25.9	28.3	18.8	16.5	24.8
POA and MOA	5.5	8.8	7.4	7.8	11.2	9.1	7.1
POA and Vowel	14.8	8.5	3.7	7.8	5.6	6.9	8.4
MOA and Vowel	0.9	2.0	3.7	1.5	4.5	3.5	1.3
POA, MOA and Vowel	2.1	1.0	2.1	0.0	3.5	1.7	1.3

(b)Distribution of errors in the performance of classifiers on the speaker independent test data set of 617 patterns.

	OCON		ACON				
Source of Errors	MLP (294)	TDNN (359)	MLP1 (235)	MLP2 (292)	TDNN1 (270)	TDNN2 (306)	DHMM (295)
POA only	38.5	32.6	33.2	37.6	37.7	39.9	41.4
MOA only	17.3	2.5	16.2	17.5	8.1	6.2	7.5
Vowel only	14.6	25.9	19.6	14.4	14.4	16.6	18.3
POA and MOA	11.2	24.8	8.5	12.0	21.1	23.2	17.6
POA and Vowel	9.2	8.4	9.8	12.3	9.5	7.2	11.2
MOA and Vowel	1.7	1.4	2.5	3.1	3.3	2.3	0.6
POA, MOA and Vowel	7.5	4.4	10.2	3.1	5.9	4.6	3.4

3.7 Summary and Conclusions

A summary of the major results of the studies presented in this chapter is given in Table-3.8.

Table 3.8: Summary of the major results of the studies on classification of the SCV segments excised from continuous speech.

1. An approach has been developed for classification of varying duration SCV segments using neural network classifiers that can handle fixed duration patterns. The patterns are derived by processing a fixed duration signal around the vowel onset points in the SCV segments.
2. The performance of classifiers based on OCON and ACON architectures, and using MLP, TDNN and DHMM models has been compared for frequently occurring SCV classes.
3. The classifiers using MLP model gave a better performance compared to TDNN and DHMM models.
4. ACON classifiers gave a better performance compared to OCON classifiers. This is mainly due to discriminatory training of ACON classifiers.
5. The performance of the ACON classifiers is more significantly affected than that of the OCON classifiers when additional classes are included.
6. A significant percentage of errors in classification of segments is due to misclassification of the place of articulation of stop consonants.

In this chapter, we have presented an approach for classification of varying duration Stop Consonant-Vowel (SCV) segments excised from continuous speech. A fixed duration signal around the vowel onset point in an SCV segment is processed to derive a pattern that is given as input to neural network classifiers. The performance of different models and architectures was evaluated for frequently occurring SCV classes. The classifiers using multilayer perceptron models give a better classification performance compared to the classifiers using TDNN models. It was also observed that the

All-Class-One-Network (ACON) classifiers show a better discriminatory capability than the One-Class-One-Network (OCON) classifiers for a given set of classes.

It was observed that the performance of all the classifiers for the speaker independent test data is poorer than the performance for the multispeaker test data. It is necessary to use parametric representations that are invariant to the variations in the characteristics of the SCV segments due to different speakers. It is also necessary to train with data from large number of speakers and use speaker adaptation techniques.

Our studies have indicated the need for appropriate parametric representations for SCV segments to obtain an improved performance. One method is to use different parametric representations appropriate for different events in the SCV segments. This method requires an approximate segmentation into regions corresponding to different significant production events, and extraction of suitable parameters from each region. It is straightforward to use patterns formed from multiple parametric representations as input to multilayer perceptron models. However multi-stream input based approaches [79] may have to be explored for TDNN and HMM models.

Mere parametric representation extracted from speech signal does not help to improve the classification accuracy beyond a limit. The parametric representation is currently viewed as a vector of data. But to take care of the variability for different repetitions and due to different speakers, it is necessary to capture the features present in the parameter data and use these features for classification. It may be possible to use neural network models to capture the features from data.

In this chapter, we have presented the studies on classification of segments belonging to a subset of SCV classes. In the next chapter, we explore suitable models for classification of the utterances of all the SCV classes. When the number of classes is large, both the OCON and ACON architectures have limitations. The discriminatory capability will be poor for OCON architectures. The limitation of the ACON architecture is that it is not possible to train a single network for large number of classes. In the next chapter we consider modular neural network architectures to handle the large number of SCV classes in Indian languages.

While the classification studies on manually excised segments corresponding to the subword units highlight the issues in parametric representation and classification, the main task is to spot these segments in continuous speech. Speech recognition by humans also takes place by spotting key segments. Therefore it is essential to develop approaches for spotting the subword units in continuous speech signal in order to realise a vocabulary independent continuous speech recognition system. In Chapter 6, we present our studies on spotting SCVs in continuous speech.

effects in continuous speech is better handled by using suitable parametric representations than trying to train the models to perform classification that is invariant to these effects.

In order to use the classifiers trained with the isolated utterance data for spotting SCV segments in continuous speech, it is necessary to address some issues arising out of the differences in the durations of the isolated utterances and the SCV segments in continuous speech. It is observed that the isolated utterances of a class are of much longer duration compared to the segments of the same class in continuous speech. In the approach presented in the previous chapter, speech signal with a duration of **100** ms around the vowel onset point in an SCV segment is processed to derive a fixed duration pattern that is input to neural network classifiers. This fixed duration pattern has most of the necessary information for classification of SCV segments in continuous speech. Because of the longer durations of the isolated utterances, it is necessary to process the speech signal with a duration of more than **100** ms around the vowel onset point. In developing classifiers for all the SCV classes that can be used for spotting, we consider different durations and study their effect on the performance of classifiers.

This chapter is organized as follows. The need for modular approaches to handle the large number of SCV classes is explained in section 4.2. Important issues in developing modular approaches are discussed in section 4.3. Specific issues related to modular approaches for classification of SCVs are presented in section 4.4. The issues in using the isolated utterance data are discussed in section 4.5. Studies on classification of all the SCV classes are presented in section 4.6.

4.2 Need for Modular Approaches

When the number of classes is large and the similarity amongst the classes is high, it is difficult to train a monolithic neural network classifier based on the All-Class-One-Network (ACON) architecture to form the necessary decision surfaces in the input pattern space. An attempt has been made to train a multilayer perceptron network for all the 80 SCV classes. It was observed that even after a large number of epochs, the total sum of squares error remained high and it did not change from one epoch to another. It shows that a single network could not be trained for large number of classes. It is possible to develop a classifier based on the One-Class-One-Network (OCON). architecture in which a separate network is trained for each class. This approach requires a large number of networks. In addition, the studies presented in the previous chapter have shown that the discriminatory capability of the OCON classifiers is poor.

Modular approaches can be used to overcome the limitations of the ACON and OCON architectures. In these approaches modularity is viewed as a manifestation of the principle of divide and conquer, which permits one to solve a complex computational task by dividing it into simpler subtasks and then combining their individual solutions [40]. In modular approaches for classification, the large number of classes are grouped into small subgroups and a separate neural network is trained for each subgroup. In the next section, we discuss the main issues in developing classifiers based on modular approach.

4.3 Issues in Modular Approaches for Classification

Two commonly used neural networks based on modular approaches are: (1) modular networks [40] [41] and (2) hierarchical networks [42]. A neural network is said to be modular if the computation performed by the network can be decomposed into two or more modules (subnetworks or subnets). Modular networks use a postprocessor to combine the outputs of the subnets. The structure of a modular network is shown in Fig.4.1.

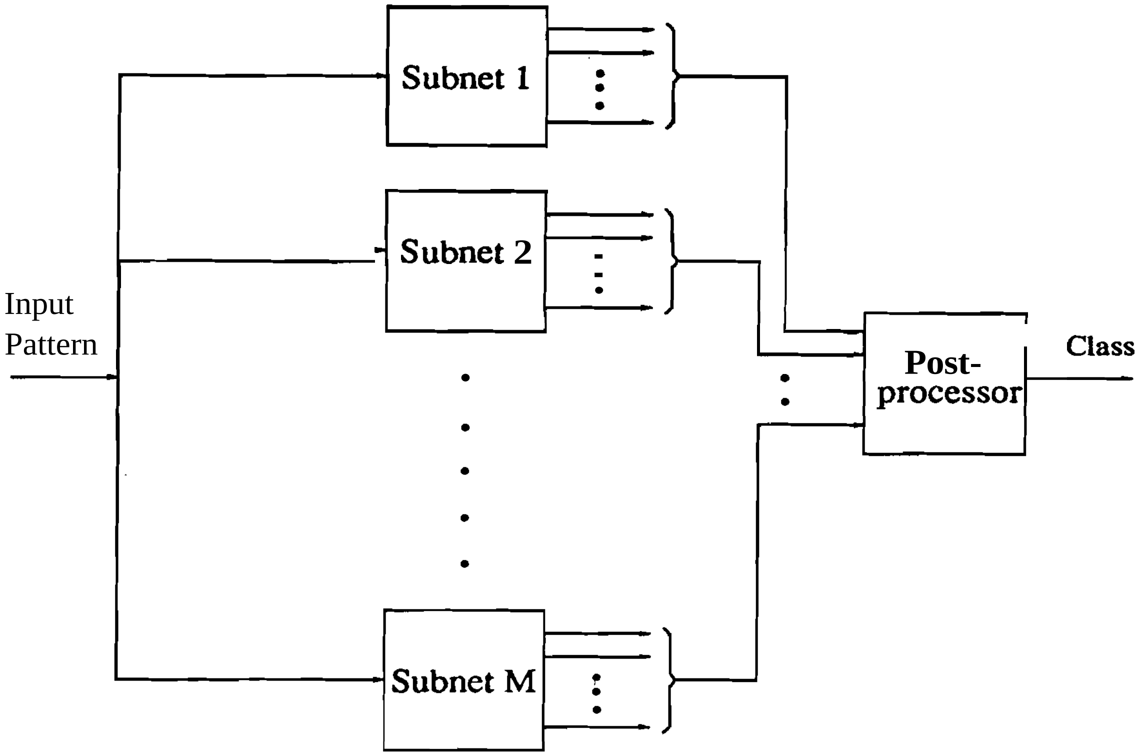


Figure 4.1: Modular network architecture for classification of patterns. The total number of classes are divided into subgroups and a separate network (subnet) is trained for each subgroup. The number of outputs of a subnet is same as the number of classes in its subgroup. For classification of a pattern, the pattern is input to each subnet. The outputs from all the subnets are combined by a postprocessor that implements the classification criterion to determine the class.

In hierarchical networks a selector network is used to determine the subgroup to which a given input pattern belongs to. Then the network for that subgroup processes the input pattern to determine its class. The structure of a hierarchical network is shown in Fig.4.2.

The main issues common to both the approaches are: (1) Selection of a criterion for

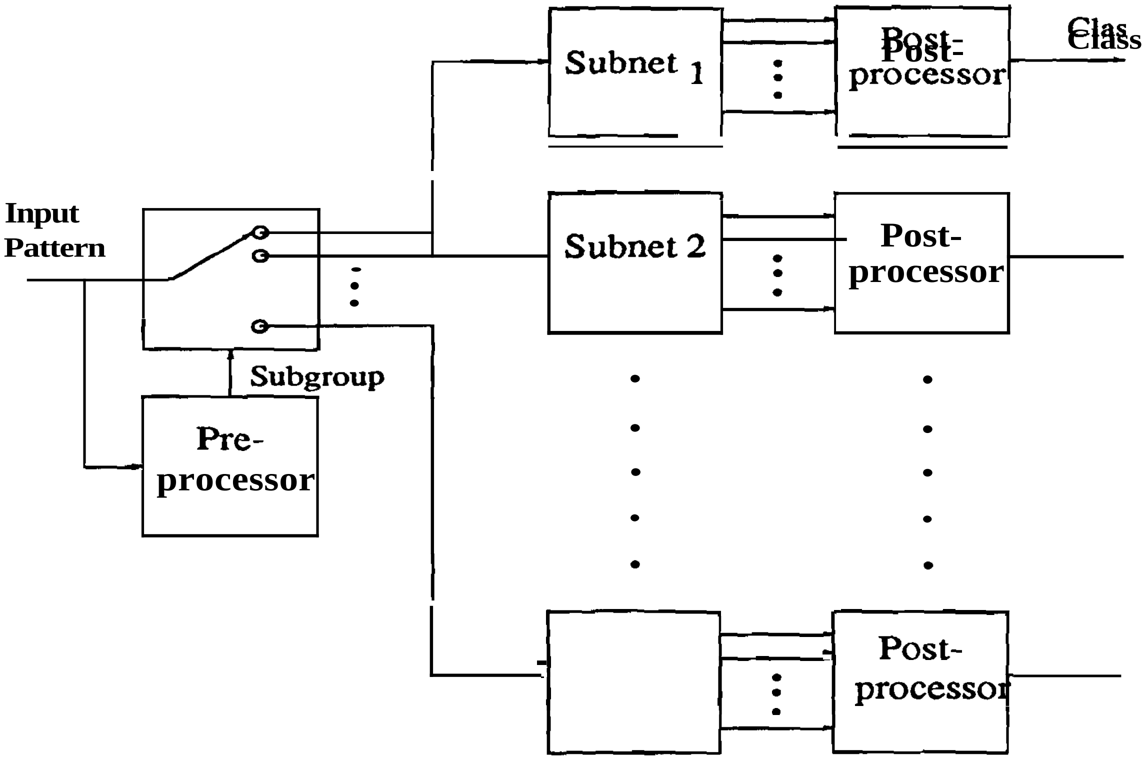


Figure 4.2: Hierarchical network architecture for classification of patterns. The total number of classes are divided into subgroups and a separate network (subnet) is trained for each subgroup. For classification of a pattern, a preprocessor determines the subgroup to which the pattern may belong to, and then the pattern is given as input to the subnet of that subgroup only. The outputs from that subnet are combined by a postprocessor to determine the class.

grouping classes into subgroups, (2) Methods for discriminatory training of subnets and (3) Choice of classification model for subnets. An issue specific to modular networks is the design of a suitable postprocessor. An issue specific to hierarchical networks is the design of a selector network. The design of the selector network is dependent on the grouping criterion chosen. The performance of a hierarchical network is critically dependent on the performance of the selector network. In our studies, we explore grouping criteria and classification models for developing modular networks for all the SCV classes.

4.4 Modular Networks for All the SCV Classes

In this section, we first consider different criteria that can be used for grouping SCV classes into subgroups. Then we discuss the issues in training the subnets. Methods for processing the outputs of the subnets are discussed in the final subsection.

4.4.1 Criteria for Grouping SCV Classes into Subgroups

The criterion used for grouping the large number of classes into subgroups decides the constitution of each subgroup. One can randomly group the classes into an arbitrarily chosen number of subgroups. Then it will be necessary to explore different number of subgroups and different ways of grouping. Instead of arbitrary grouping, we consider criteria guided by the phonetic descriptions of the SCV classes. Such criteria are useful in analyzing the performance of the classifiers and determining the sources of errors in classification.

A unique phonetic description can be given for each of the 80 SCV classes in terms of three features, namely, (1) the manner of articulation (MOA) of the stop consonant in that SCV, (2) the place of articulation (POA) of the stop consonant and (3) the identity of the vowel in that SCV. For example, the class /ka/ is described as 'Unvoiced unaspirated velar stop consonant followed by the vowel /a/'. The phonetic description of the SCV classes suggests that grouping can be done in such a way that one of the three features is common to the classes in a subgroup. This results in three criteria that can be considered for grouping.

Grouping based on MOA leads to four subgroups: (1) Unvoiced-Unaspirated (UVUA), (2) Unvoiced-Aspirated (UVA), (3) Voiced-Unaspirated (VUA) and (4) Voiced-Aspirated (VA). Each subgroup consists of 20 classes and the stop consonants in these classes have the same manner of articulation. The classes in each of these subgroups are given in Table-4.1(a).

Table 4.1: Classes in subgroups based on different grouping criteria.

(a) Classes in MOA subgroups

MOA	SCV Classes				
UVUA	ka	ki	ku	ke	ko
	ṭa	ṭi	ṭu	ṭe	ṭo
	ta	ti	tu	te	to
	pa	pi	pu	pe	po
UVA	kha	khi	khu	khe	kho
	tha	ṭhi	ṭhu	the	tho
	tha	thi	thu	the	tho
	pha	phi	phu	phe	pho
VUA	ga	gi	gu	ge	go
	da	di	du	de	do
	da	di	du	de	do
	ba	bi	bu	be	bo
VA	gha	ghi	ghu	ghe	gho
	dha	dhi	dhu	dhe	dho
	dha	dhi	dhu	dhe	dho
	bha	bhi	bhu	bhe	bho

(b) Classes in POA subgroups

POA	SCV Classes				
Velar	ka	ki	ku	ke	ko
	kha	khi	khu	khe	kho
	ga	gi	gu	ge	go
	gha	ghi	ghu	ghe	gho
Alveolar	fa	fi	ṭu	ṭe	ṭo
	ṭha	ṭhi	ṭhu	the	tho
	da	ḍi	du	de	do
	dha	ḍhi	dhu	dhe	dho
Dental	ta	ti	tu	te	to
	tha	thi	thu	the	tho
	da	di	du	de	do
	dha	dhi	dhu	dhe	dho
Bilabial	pa	pi	pu	pe	po
	pha	phi	phu	phe	pho
	ba	bi	bu	be	bo
	bha	bhi	bhu	bhe	bho

(c) Classes in Vowel subgroups

Vowel	SCV Classes			
/a/	ka	kha	ga	gha
	ṭa	ṭha	da	dha
	ta	tha	da	dha
	pa	pha	ba	bha
/i/	ki	khi	gi	ghi
	ṭi	thi	di	dhi
	ti	thi	di	dhi
	pi	phi	bi	bhi
/u/	ku	khu	gu	ghu
	ṭu	ṭhu	du	dhu
	tu	thu	du	dhu
	pu	phu	bu	bhu
/e/	ke	khe	ge	ghe
	ṭe	the	de	dhe
	te	the	de	dhe
	pe	phe	be	bhe
/o/	ko	kho	go	gho
	ṭo	ṭho	do	dho
	to	tho	do	dho
	po	pho	bo	bho

Grouping based on POA leads to four subgroups: (1) Velar, (2) Alveolar, (3) Dental and (4) Bilabial. Each subgroup consists of 20 classes and the stop consonants in these classes have the same place of articulation. The classes in each of these subgroups are given in Table-4.1(b).

Grouping based on the vowel in SCVs leads to five subgroups with one subgroup for each of the five vowels: /a/, /i/, /u/, /e/ and /o/. Each subgroup consists of 16 classes and these classes have the same vowel. The classes in each of these subgroups are given in Table-4.1(c).

We consider each of the three grouping criteria in developing a modular network for all the SCV classes. The classification performance of a modular network based on a particular grouping depends on the performance of its subnets. The performance of subnets is dependent on the data used for training them. We next discuss the issues in training the subnets.

4.4.2 Training of Subnets

The training data set for a subnet should generally consist of patterns belonging to the classes in its subgroup only. Then each subnet is trained as an ACON classifier to form the decision surfaces for the classes in its subgroup. When a modular network is used for classification, a given test pattern is input to all its subnets and the outputs of the subnets are processed to determine the class. In order to correctly classify, the output value for the class of the test pattern should be high and all other output values should be low. It may be necessary to train each subnet with a few patterns belonging to the classes of the other subnets. These patterns can be considered as negative examples. For a negative example pattern, the subnet should be trained to give a low value for all its outputs. The aim of using negative examples in training a subnet is to form a decision boundary around the regions of its classes.

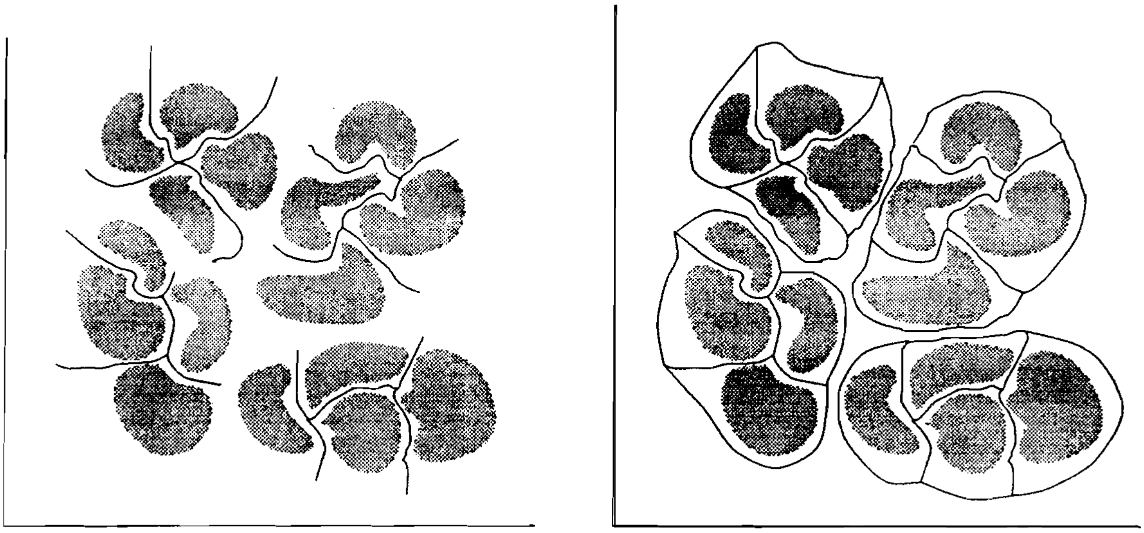
The effect of including negative **examples** in the training data sets of **subnets** is illustrated in Fig.4.3 for an arbitrary **2-dimensional** pattern space. Typical decision surfaces expected to be formed by **subnets** are shown in **Fig.4.3(a)**. Here we consider 16 classes that are divided into 4 subgroups with 4 classes in each subgroup. Each **subnet** is separately trained with patterns belonging to the classes in its subgroup only. Therefore for each **subnet,** the decision surfaces are formed among the regions of classes in the subgroup only. When the patterns of the classes in the other subgroups are included as negative examples, it is expected that a boundary be formed around the regions of the classes in its subgroup. The expected effect of including negative examples is **shown** in **Fig.4.3(b)**. In our studies, we compare the performance of the modular networks with and without negative examples being used in training the **subnets**.

4.4.3 Processing the Outputs of Subnets

A simple way of processing the outputs of **subnets** is to assign the class with the largest value among the outputs of all the **subnets**. Because of the similarity amongst the classes with in a subgroup and **also** amongst several classes in different subgroups, we use a method in which the classes with the largest and the second largest output values of each **subnet** are also considered in deciding the class.

4.5 Derivation of Patterns for Isolated Utterance Data

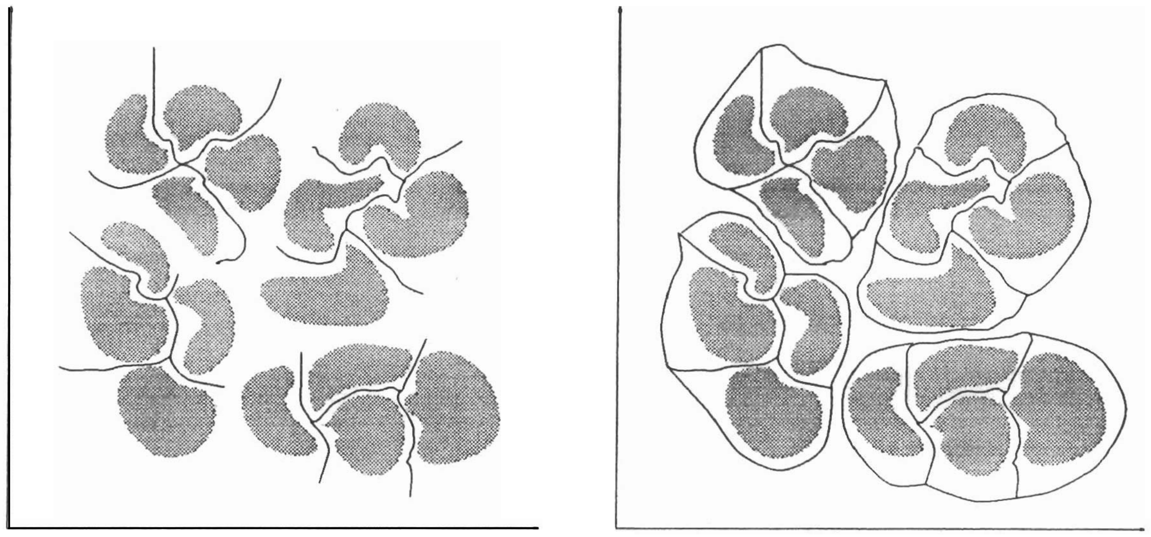
The approach proposed in the previous chapter for classification of SCV segments in continuous speech can be extended for the isolated utterance data. In this approach, a fixed duration portion of the signal around the Vowel Onset Point (VOP) of an SCV utterance is processed to derive a pattern. There are two important factors that have to be taken into consideration for deciding the duration of the portion of the signal to



(a) Training subnets without negative examples

(b) Training subnets with negative examples

Figure **4.3:** Illustration of decision surfaces and boundaries formed by subnet classifiers in modular networks. An arbitrary 2-dimensional pattern space is used to explain the effect of including negative examples in training data sets of subnets. Typical decision surfaces expected to be formed by subnets are shown in (a). Here 16 classes are divided into 4 subgroups with 4 classes in each subgroup. Each subnet is separately trained with patterns belonging to the classes in its subgroup only. Therefore for each subnet, the decision surfaces are formed among the regions of classes in the subgroup only. When the patterns of the classes in the other subgroups are included as negative examples in the training data set of a subnet, it is expected that a boundary be formed around the regions of the classes in its subgroup. The expected effect of including negative examples is shown in (b).



(a) Training subnets without negative examples

(b) Training subnets with negative examples

Figure 4.3: Illustration of decision surfaces and boundaries formed by subnet classifiers in modular networks. An arbitrary 2-dimensional pattern space is used to explain the effect of including negative examples in training data sets of subnets. Typical decision surfaces expected to be formed by subnets are shown in (a). Here 16 classes are divided into 4 subgroups with 4 classes in each subgroup. Each subnet is separately trained with patterns belonging to the classes in its subgroup only. Therefore for each subnet, the decision surfaces are formed among the regions of classes in the subgroup only. When the patterns of the classes in the other subgroups are included as negative examples in the training data set of a subnet, it is expected that a boundary be formed around the regions of the classes in its subgroup. The expected effect of including negative examples is shown in (b).

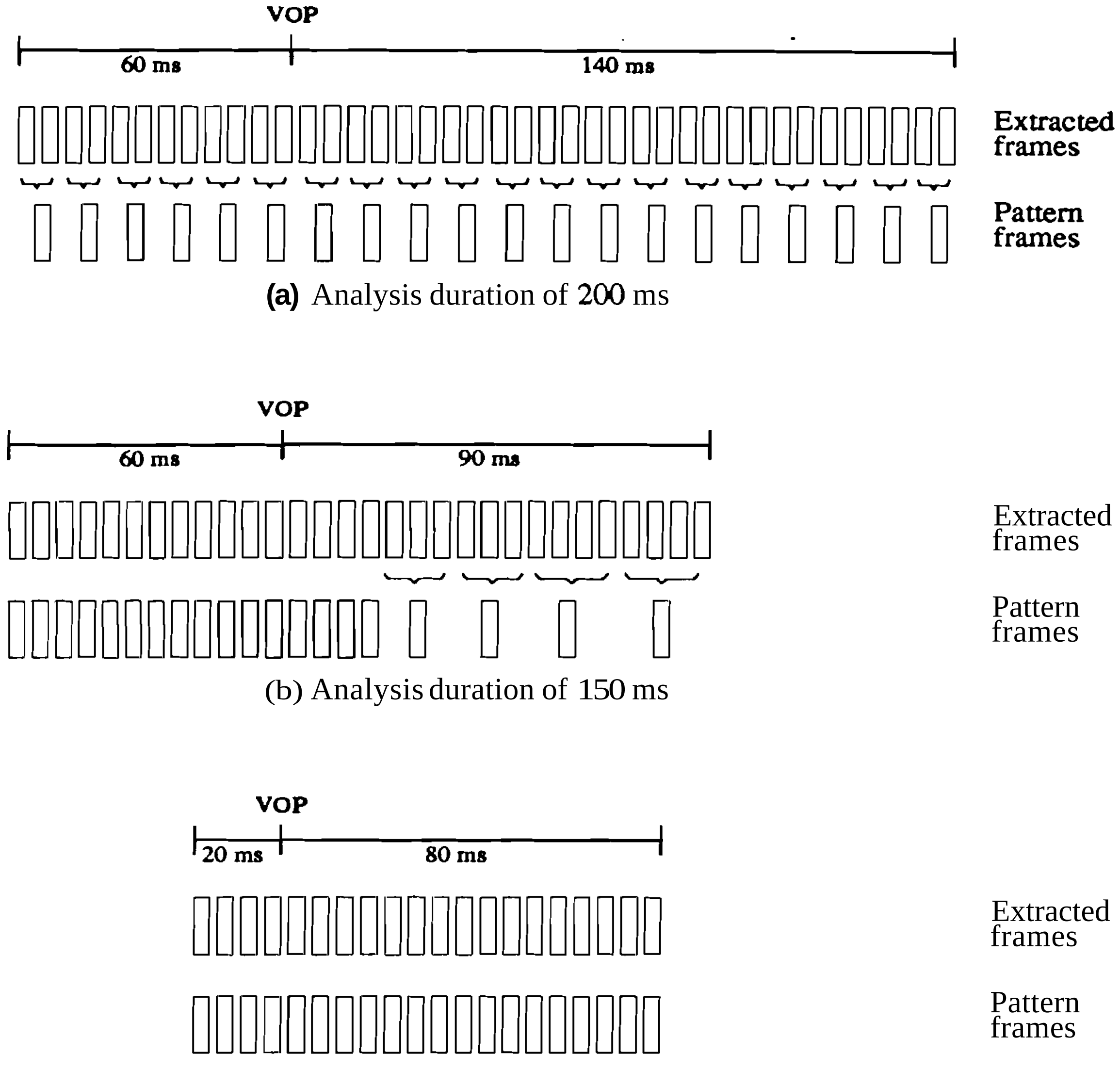
be processed for the isolated utterance data. The first factor is that the portion of the signal should have all the necessary information for classification. The second factor is that it should be possible to use the networks trained with the isolated utterance data for spotting SCV segments in continuous speech. For this, the patterns derived from the isolated utterance data have to be matched with the patterns derived from continuous speech segments. We consider three different durations in our studies.

We first consider the method for deriving a pattern from a 200 ms portion of the signal with 60 ms before and 140 ms after the VOP. This fixed duration signal is processed to extract 40 frames with 12 weighted cepstral coefficients in each frame. In order to reduce the size of the pattern, the average coefficients for every two adjacent frames are used. Thus a 20 frame pattern is used to represent an SCV utterance. This method for derivation of patterns for the segment duration of 200 ms is shown in Fig.4.4(a).

We next present the method used for deriving a pattern from a 150 ms portion of the signal with 60 ms before and 90 ms after the VOP. The signal is processed to extract 30 frames of weighted cepstral coefficients. The first 16 frames are retained and the remaining 14 frames are averaged to get the other 4 frames in the 20 frame pattern, as shown in Fig.4.4(b).

Finally we consider a 100 ms portion of the signal with 20 ms before and SO ms after the VOP. This signal is processed to extract 20 frames using a frame size of 20 ms and a shift of 5 ms. All the 20 frames are used as the pattern frames, as shown in Fig.4.4(c).

In our studies, we compare the performance of subnets and modular networks for different segment durations.



(a) Analysis duration of 200 ms

(b) Analysis duration of 150 ms

(c) Analysis duration of 100 ms

Figure **4.4:** Derivation of fixed duration patterns from speech signal of SCVs using different segment durations. The method used for derivation of a pattern of **20** frames is shown for the segment duration of (a) **200** ms, (b) **150** ms, and (c) **100** ms. The portions of the speech signal around the vowel onset point (VOP) processed, the frames extracted, and the pattern frames derived from the extracted frames are indicated for each duration. For the duration of **200** ms, **40** frames are extracted and the adjacent frames are averaged. For the duration of **150** ms, **30** frames are extracted. The first **16** frames are retained and the other **14** frames are averaged as shown to obtain the remaining **4** frames of the pattern. For the segment duration of **100** ms, **20** frames extracted from speech signal are used **as** the pattern frames.

4.6 Classification Studies and Results

4.6.1 Implementation details

Isolated utterance data for the 80 SCV classes was collected from three male speakers. For each class, 12 tokens were collected from each speaker. In the studies presented in this chapter, the training data for a class includes four tokens from each speaker. The remaining eight tokens from each speaker for a class are used as the test data. A pattern consisting of 20 frames with 12 weighted cepstral coefficients per frame is derived by processing the speech signal with a particular duration around the vowel onset point. The vowel onset points in the SCV utterances are detected using the method presented in the previous chapter.

We consider the multilayer perceptron (MLP), time-delay neural network (TDNN) and discrete hidden Markov model (DHMM) to build the subnets. The MLP model has 70 nodes in the first hidden layer and 50 nodes in the second hidden layer. The TDNN model has a single hidden layer with 20 nodes. The DHMM is a 5-state, left-to-right model.

In our first set of studies, we evaluate the performance of subnets based on different grouping criteria and built using different models. Patterns derived using a segment duration of 200 ms are used in these studies. In the second set of studies, we compare the performance of subnets for different segment durations. Finally we study the performance of modular networks. In the remaining part of this section, we describe· these studies and present the results. The studies carried out in this chapter are listed in Table-4.2.

4.6.2 Performance of Subnets using Different Models

The aim of the studies presented in this subsection is to evaluate and analyze the classification performance of subnets based on different grouping criteria and built

Table 4.2: List of studies on development of modular networks for classification of large number of SCV classes.

1. Comparison of the performance of **subnets** for subgroups of SCV classes formed using different grouping criteria.
2. Comparison of the performance of **subnets** using MLP, TDNN and DHMM models.
3. Analysis of the performance of **subnets** based on different grouping criteria to identify the sources of errors in classification.
4. Comparison of the performance of **subnets** based on different grouping criteria for each SCV class.
5. Comparison and analysis of the performance of **subnets** for different segment durations.
6. Performance of modular networks based on different grouping criteria and for different segment durations.
7. Comparison of the performance of modular networks using different data sets for training the **subnets**.
8. Comparison of the performance of modular networks based on different grouping criteria for each SCV class.

using different classification models. In these studies, we consider patterns derived from 200 ms long segments. The training and the test data sets of a **subnet** includes the patterns belonging to the classes in its subgroup only. The performance of **subnets** for different subgroups is given in **Table-4.3**. The performance is given a percentage of the total number of patterns in a data set that are correctly classified by a **subnet**.

Grouping based on POA gives the best average performance irrespective of the model used. The average performance for the other two groupings is approximately same. The better performance for grouping based on POA can be explained as follows. The **subnets** for POA subgroups have to be trained to discriminate only the manners

Table 4.3: Classification performance of **subnets** for subgroups of SCV classes formed **using** different grouping criteria. The performance for **subnets** based on MOA grouping is given in (a), for **subnets** based on POA grouping in (b) and for **subnets** based on vowel grouping in (c). The performance is given for different models. **Subnets** based on POA grouping give the best average performance. **Subnets** built using MLP models give a better performance compared to the other models.

(a) Performance of **subnets** based on MOA grouping.

MOA	Training Data			Test Data		
Subgroup	MLP	TDNN	DHMM	MLP	TDNN	DHMM
UVUA	98.1	80.0	97.9	77.1	65.0	57.9
UVA	98.1	79.8	95.8	70.0	58.8	46.3
VUA	94.2	67.5	86.0	62.9	50.4	45.8
VA	91.0	71.0	86.9	56.3	47.5	37.5
Average	95.4	74.2	91.7	66.6	54.4	46.9

(b) Performance of **subnets** based on POA grouping.

POA	Training Data			Test Data		
Subgroup	MLP	TDNN	DHMM	MLP	TDNN	DHMM
Velar	95.6	74.2	91.5	54.6	46.7	48.3
Alveolar	95.0	80.6	96.0	84.2	64.6	66.3
Dental	93.3	71.9	94.4	77.9	56.7	55.8
Bilabial	91.7	76.0	92.5	80.0	58.8	58.8
Average	93.9	75.7	93.6	74.2	56.6	57.4

(a) Performance of **subnets** based on Vowel grouping.

Vowel	Training Data			Test Data		
Subgroup	MLP	TDNN	DHMM	MLP	TDNN	DHMM
/a/	92.7	82.0	95.1	61.5	60.4	50.0
/i/	93.2	67.2	92.0	66.6	42.7	40.1
/u/	92.2	67.4	89.3	68.8	52.6	41.7
/e/	94.0	79.9	89.3	62.0	55.7	36.5
/o/	91.7	77.1	90.6	67.7	56.8	45.3
Average	92.8	74.7	91.3	65.1	53.6	42.7

of articulation and the vowels. The **subnets** for the other groupings have to be trained to discriminate the places of articulation. It is more difficult to discriminate the places of articulation because the necessary information is mainly present in the transition regions with dynamic spectral characteristics.

An analysis of the performance of **subnets** was carried out to determine the distribution of errors due to misclassification of each of the following: (1) POA only, (2) MOA only, (3) vowel only, (4) POA and MOA, (5) POA and vowel, (6) MOA and vowel, and (7) POA, MOA and vowel. The distribution of errors for the MLP based **subnets** is given in Table-4.4.

Table 4.4: Distribution of errors in classification performance of MLP based **subnets** for different grouping criteria, due to **misclassification** of one or more of the following three features of an SCV: (1) Manner of articulation (MOA) of the consonant, (2) Place of articulation (POA) of the consonant, and (3) Vowel. For a particular grouping criterion, the SCV classes in a subgroup have one of these features common to all of them. Therefore the errors in the performance of **subnets** based on a particular criterion can be due to misclassification of the other two features only. It can be noted that for the criteria of MOA and Vowel, more than 50% of the errors are due to misclassification of POA feature only.

Grouping Criterion	Source of Errors	Percentage of Errors
MOA	POA only	55.3
	Vowel only	22.7
	POA and Vowel	22.0
POA	MOA only	39.9
	Vowel only	35.1
	MOA and Vowel	25.0
Vowel	MOA only	22.7
	POA only	50.1
	MOA and POA	27.2

It can be noted that about 50% of the total number of errors in the performance for the grouping criteria of MOA and Vowel are due to misclassification of the POA only. This analysis supports the explanation given above.

It is observed that the subnets for the unaspirated SCV classes (UVUA and VUA subgroups) give a better performance than the corresponding subnets for the aspirated SCV classes (UVA and VA subgroups). The poorer performance for the aspirated SCVs can be due to absence of the necessary discriminatory information in the patterns derived. The beginning of the aspiration region may have been identified as the vowel onset point. It is necessary to evolve a better technique for deriving the patterns from the aspirated SCV utterance data.

The results of the studies indicate that the subnets using multilayer perceptron model gave a better performance on the test data compared to TDNN and DHMM models. The poorer performance of the TDNN model can be due to the small number of hidden nodes used. This is indicated by its poor performance for the training data. It may be necessary to increase the number of hidden nodes to improve the performance. The TDNN models with large number of hidden nodes require long training periods and large training sets. The better generalization capability of the MLP model compared to DHMM is indicated by the difference in their performance on the test data though their performance on the training data is approximately same.

We examine the performance of subnets on the test data when the class with the second largest output value is also considered in deciding the class. The average performance of subnets for different grouping criteria is given in Table-4.5 for two cases of deciding the class: (1) Correct class is the class with the largest value among the outputs of a subnet (Case-1) and (2) Correct class is amongst the classes with the largest and the second largest output values of a subnet(Case_2). It can be noted that there is an increase of about 12 to 22% in the performance for Case_2 over Case_1 indicating the similarity even amongst the classes in a subgroup.

The classification performance of a subnet for patterns belonging to an SCV class

Table 4.5: Average performance of subnets based for the two cases of classification criterion: (1) Case-1 that an input pattern is correctly classified if its class is the class with the largest value amongst the outputs of a subnet, and (2) Case2 that an input pattern is correctly classified if its class is amongst the classes with the largest and the second largest output values of a subnet. The increase in the performance for Case2 over Case_1 indicates that for many patterns that are incorrectly classified by a subnet, the class with the second largest value is the correct class.

Grouping	Case_1			Case_2		
Criterion	MLP	TDNN	DHMM	MLP	TDNN	DHMM
MOA	66.6	54.4	46.9	79.5	72.2	65.5
POA	74.2	56.6	57.4	86.0	78.5	74.8
Vowel	65.1	53.6	42.7	78.7	67.8	60.2

depends not only on the characteristics of these patterns but also on the characteristics of the patterns belonging to the other classes in the subgroup. The shape of the decision surface for a class depends on the other classes for which a subnet is trained. Therefore the performance of subnets on the test data of a class can vary for different grouping criteria. For example, the class /ka/ belongs to the UVUA subgroup for grouping based on MOA, to the 'Velar' subgroup for grouping based on POA and to the '/a/' subgroup for grouping based on vowel. The constitution of each of these subgroups is different and hence the decision surface for the class /ka/ will be of differentshape in their subnets. The performance of subnets for each of the classes in their subgroups is given in Table-4.6. Each entry in the table shows the percentage of the total number of test patterns of an SCV class that are correctly classified by its subnet. Here we consider the performance of subnets built using the MLP model. It can be noted from Table-4.6 that subnets of the three grouping criteriado not give the same performance for many classes.

Table 4.6: Classification performance (in percentage) of **subnets** based on different grouping criteria for test data of each SCV class.

SCV	Grouping Criterion			SCV	Grouping Criterion		
Class	MOA	POA	Vowel	Class	MOA	POA	Vowel
ka	58	83	42	**ṭa**	92	92	83
kha	92	67	92	**ṭha**	92	83	58
ga	67	83	. 83	da	83	67	67
gha	50	50	25	dha	67	75	58
ki	58	50	58	**ṭi**	75	83	58
khi	50	67	58	**ṭhi**	67	100	58
gi	42	33	67	di	67	67	42
ghi	33	42	50	dhi	67	67	58
ku	67	75	83	**ṭu**	83	100	92
khu	75	58	67	**ṭhu**	58	92	67
gu	50	50	25	du	67	92	75
ghu	33	17	33	dhu	58	67	58
ke	75	67	50	**ṭe**	92	92	100
khe	58	42	42	the	67	92	75
ge	50	42	67	de	50	92	58
ghe	58	42	42	dhe	92	92	67
ko	75	67	92	**ṭo**	75	83	75
kho	67	58	75	tho	75	92	50
go	42	42	67	do	83	83	83
gho	67	58	17	dho	75	75	67
ta	100	100	100	pa	92	92	75
tha	42	58	42	pha	92	100	58
da	67	67	42	ba	83	83	75
dha	67	92	50	bha	67	75	33
ti	92	75	67	pi	50	42	67
thi	75	100	92	phi	92	83	92
di	75	S3	58	bi	75	75	75
dhi	58	83	42	bhi	42	100	92
tu	75	83	67	pu	83	67	83
thu	67	67	50	phu	42	67	75
du	75	42	75	bu	92	83	75
dhu	42	75	92	bhu	75	58	83
te	75	92	83	pe	83	100	67
the	75	100	58	phe	67	92	83
de	42	58	8	be	58	58	33
dhe	50	100	75	bhe	75	83	83
to	67	75	67	**po**	75	92	75
tho	75	75	83	pho	75	92	92
do	42	67	58	bo	50	75	58
dho	42	67	50	bho	83	83	75

The studies presented in this subsection are concerned with evaluation and analysis of the performance of subnets based on different grouping criteria and using different classification models. Fixed duration patterns derived from the speech signal with the duration of 200 ms around the vowel onset points have been used in these studies. In the next subsection, we consider the performance of subnets using patterns derived using different segment durations.

4.6.3 Performance of Subnets for Different Segment Durations

In this subsection we study the effects of the durations used for deriving the patterns on the performance of subnets. As mentioned in section 4.5, three different durations are considered. The performance of subnets for the segment duration of 200 ms was given in the previous subsection. Similar studies have been carried out for the segment durations of 150 ms and 100 ms. These studies have been limited to the multilayer perceptron models because these models have given a better performance. The performance of subnets for different segment durations is given in Table-4.7.

The performance for subnets based on MOA grouping indicates that there is a significant decrease in the performance for aspirated SCV classes (UVA and VA subgroups) when the segment duration is reduced from 200 ms to 150 ms and 100 ms. The performance of subnets for POA subgroups shows that reduction of duration from 200 ms to 150 ms has affected the performance for subgroups other than the 'Velar' .subgroup. The decrease in performance is higher for all the subgroups when the duration is reduced to 100 ms. It is observed from the performance of subnets for vowel subgroups is not affected significantly when the duration is reduced from 200 ms to 150 ms. When the duration is reduced to 100 ms, there is a significant decrease in the performance for all subgroups with an exception of the '/a/' subgroup.

In order to determine the effects of the segment durations on the performance

Table 4.7: Classification performance of **subnets** for different segment durations. The performance for **subnets** based on MOA grouping is given in (a), for **subnets** based on POA grouping in (b) and for **subnets** based on vowel grouping in (c). The segment duration of 200 ms gives a better performance compared to the durations of 150 ms and 100 ms.

(a) Performance of **subnets** based on MOA grouping.

MOA	Training Data			Test Data		
Subgroup	**200ms**	**150ms**	**100ms**	**200ms**	**150ms**	**100ms**
UVUA	98.1	99.2	99.6	77.1	64.1	62.7
UVA	98.1	98.8	97.9	70.0	40.4	31.0
VUA	94.2	99.6	100.0	62.9	65.8	64.0
VA	91.0	99.2	98.8	56.3	43.8	35.0
Average	95.4	99.2	99.1	66.6	53.5	48.2

(a) Performance of **subnets** based on POA grouping.

POA	Training Data			Test Data		
Subgroup	**200ms**	**150ms**	**100ms**	**200ms**	**150ms**	**100ms**
Velar	95.6	99.6	97.5	54.6	60.8	**51.5**
Alveolar	95.0	99.6	98.8	84.2	62.5	54.6
Dental	93.3	99.6	96.7	77.9	68.1	51.0
Bilabial	91.7	99.6	98.3.	80.0	71.5	56.5
Average	93.9	99.6	97.8	74.2	65.7	53.4

(a) Performance of **subnets** based on POA grouping.

Vowel	Training Data			Test Data		
Subgroup	**200ms**	**150ms**	**100ms**	**200ms**	**150ms**	**100ms**
/a/	92.7	97.4	98.4	61.5	69.8	64.1
/i/	93.2	99.5	94.3	66.6	63.5	31.3
/u/	92.2	100.0	97.9	68.8	63.0	45.8
/e/	94.0	99.0	96.4	62.0	61.5	41.7
/o/	91.7	100.0	95.8	67.7	69.5	40.4
Average	92.8	99.2	96.6	65.1	65.5	**44.6**

of **subnets**, the distribution of errors in classification is obtained. The distributions for different grouping criteria are given in Table-4.8. This table gives the number of errors in classification performance of a **subnet** due to misclassification of one or more of the three features, namely, MOA, POA and vowel.

Table 4.8: Distribution of errors in classification performance of **subnets** for different segment durations.

Grouping Criterion	Source of Errors	Number of Errors		
		200ms	**150ms**	**100ms**
MOA	POA only	352	394	515
	Vowel only	146	280	217
	POA and Vowel	142	218	263
POA	MOA only	198	209	465
	Vowel only	174	349	247
	MOA and Vowel	124	100	188
Vowel	MOA only	152	174	294
	POA only	336	401	414
	MOA and POA	182	88	355

The distribution of errors for MOA grouping indicates that when the duration is reduced from 200 ms to 150 ms, the number of errors due to misclassification of vowel has increased significantly. It is observed that many of these errors are for aspirated SCV utterances. This indicates that the discriminatory information about vowel has been lost in derivation of patterns for aspirated SCV data. When the duration is reduced to 100 ms, the number of errors has increased due to misclassification of POA as well as vowel.

The distribution for POA grouping indicates that the number of errors due to misclassification of vowel has increased significantly when the duration is reduced to 150 ms. The number of errors due to misclassification of MOA has increased more significantly than that due to misclassification of vowel for the duration of 100 ms. This can be mainly due to the short duration of 20 ms before VOP being used in the duration of 100 ms.

The distribution of errors for vowel grouping indicates that the number of errors due to misclassification of POA only is not much different for durations of 150 ms and 100 ms. Therefore the increase in the total number of errors for duration of 100 ms is mainly due to misclassification of MOA.

The studies on performance of **subnets** for different segment durations have shown that the duration of 200 ms gives the best average **performance** irrespective of the grouping criterion. The decrease in the performance for the duration of 150 ms is mainly due to misclassification of vowel. The decrease in the performance for the duration of 100 ms is found to be mainly due to misclassification of MOA and and to a lesser extent due to misclassification of POA.

The focus of the studies presented so far has been on the performance of **subnets** in classification of patterns belonging to the classes in their subgroups only. In the next subsection, we study the performance of modular networks in deciding the class of a pattern belonging to any of the 80 **SCV** classes.

4.6.4 Classification Performance of Modular Networks

In order to determine the class of a pattern belonging to any of the 80 **SCV** classes, it is necessary to process the outputs of **subnets** in the modular network based on a particular grouping criterion. The following method is used to determine the class. Let A; be the largest output value of the ith **subnet** and N be the number of **subnets** for a particular grouping criterion. Then the class with $A_{Max} = \max \{ A_i \}$, $i =$ 1,2,...,N, can be assigned to the input pattern. Because of the large number of classes and similarity amongst several classes, the correct class may be the class with the second largest output in a **subnet** or the class with the second largest value amongst the A_is. This suggests that one can also consider the classes with the output values that are close to that of A_{Max}. Let A_{i1} and A_{i2} be the largest and the second largest output values of the ith subnet. Considering the M largest values amongst the set of values { A_{i1}, A_{i2} }, i = 1,2,...,N, the performance of the modular networks can be given for different values of M. In our studies, we give the performance for M = 1, 2, 3 and 4, called as Case-1, Case_2, Case_3 and Case-4, respectively.

We consider the performance of modular networks with subnets built using multilayer perceptron model only because the subnets built using this model gave the best performance. The average performance on the test data of all the 80 SCV classes for the modular networks based on different grouping criteria and for different segments durations is given in Table-4.9.

Table 4.9: Classification performance on test data of all SCV classes for the modular networks based on different grouping criteria and for different segment durations. The performance is given for four different cases of classification criterion. The Case-M of classification criterion corresponds to the case when the class of an input pattern is amongst the classes of the M largest output values of all subnets in a modular network. The modular networks for the POA grouping give a better performance compared to other two groupings. The performance for the duration of 200 ms is better than that for the reduced durations. The significant increase in the performance for Case2 over Case-1 indicates that for many patterns that are incorrectly classified by the modular network, the class with the second largest value among 80 outputs is the correct class.

Grouping	Segment	Classification Criterion			
criterion	Duration	Case-1	Case2	Case-3	Case-4
MOA	200ms	29.2	50.2	59.0	65.3
	150ms	22.2	38.2	48.3	54.4
	100ms	19.0	31.3	40.8	47.2
POA	200ms	35.1	56.9	69.5	76.6
	150ms	30.8	49.4	60.1	66.7
	100ms	20.6	36.4	47.0	53.9
Vowel	200ms	30.1	47.5	58.8	63.6
	150ms	21.4	37.0	49.8	59.3
	100ms	13.0	24.9	33.0	37.9

A performance of about 75% is obtained for the modular network when the classes with the four largest output values amongst the 80 SCV classes are considered. This performance is significant considering the large number of classes and confusability amongst several classes. The modular networks for the POA grouping give a better performance compared to the other two groupings. It is also seen that the performance for the duration of 200 ms is better than that for the reduced durations. This behaviour in the performance of the modular networks reflects the performance of the subnets. It is important to evolve techniques to reduce the number of errors in classification at the level of subnets in order to improve the performance of the modular networks.

It is observed that the performance of the modular networks for Case-1 is much less than the average performance of the subnets. One of the reasons for this behaviour is that each of the subnets is not trained to give low output values for the patterns belonging to the classes of the other subnets. A study is carried out in which each subnet is trained to give low output values for patterns (negative examples) belonging to the classes of the other subnets. This study is carried out for the subnets built using multilayer perceptron model and for the segment duration of 200 ms. A comparison of the performance of the modular networks using subnets trained without and with the negative examples included in their training data sets is given in Table-4.10.

It can be seen that the performance for Case-1 increases by about 12 to 20% for all the grouping criteria when the .negative examples are included. A performance of about 50% can be obtained for the modular networks in classification of all the SCV classes even when only the class with the largest value amongst the outputs for 80 classes is considered.

It has been observed earlier that the performance of subnets based on different grouping criteria is not uniform for many of the SCV classes. A similar behaviour has

Table 4.10: Comparison of the performance on test data of all the SCV classes for the MLP based modular networks without and with the negative examples included in the training data sets of their subnets. The comparison of the performance is given for different groupings and for different cases of classification criterion. The performance for Case-1 of classification criterion increases significantly when the negative examples are included.

Grouping criterion	Without negative examples in training data				With negative examples in training data			
	Case-1	Case_2	Case_3	Case-4	Case-1	Case_2	Case_3	Case-4
MOA	29.2	50.2	59.0	65.3	49.8	59.7	65.6	69.8
POA	35.1	56.9	69.5	76.6	47.4	64.1	71.3	75.1
Vowel	30.1	47.5	58.8	63.6	50.6	64.6	70.0	73.3

also been observed in the performance of the modular networks. The performance on test data of each SCV class for the modular networks based on different grouping criteria is given in Table-4.11. The performance is given for **MLP** based modular networks and for the duration of 200 ms. The differences in the performances are significant for many classes. For example, the modular network based on **MOA** grouping gives a performance of 41% for the class /ṭo/. The network based on **POA** grouping gives a performance of 75% and the network based on vowel grouping gives a performance of only 16%.

It is also observed that the modular networks based on different groupings correctly classified different patterns in the test set of a class. Therefore, though the average performance of different networks for a class is same, it is not necessary that all of them correctly classified the same subset of patterns. **A** study has been carried out to determine the percentage of the total number of test patterns of all the classes that have been correctly classified by different number of networks. The results of this study are presented for different segment durations in Table-4.12.

Table **4.11:** Classification performance (in percentage) of modular networks based on different grouping criteria for test data of each SCV class.

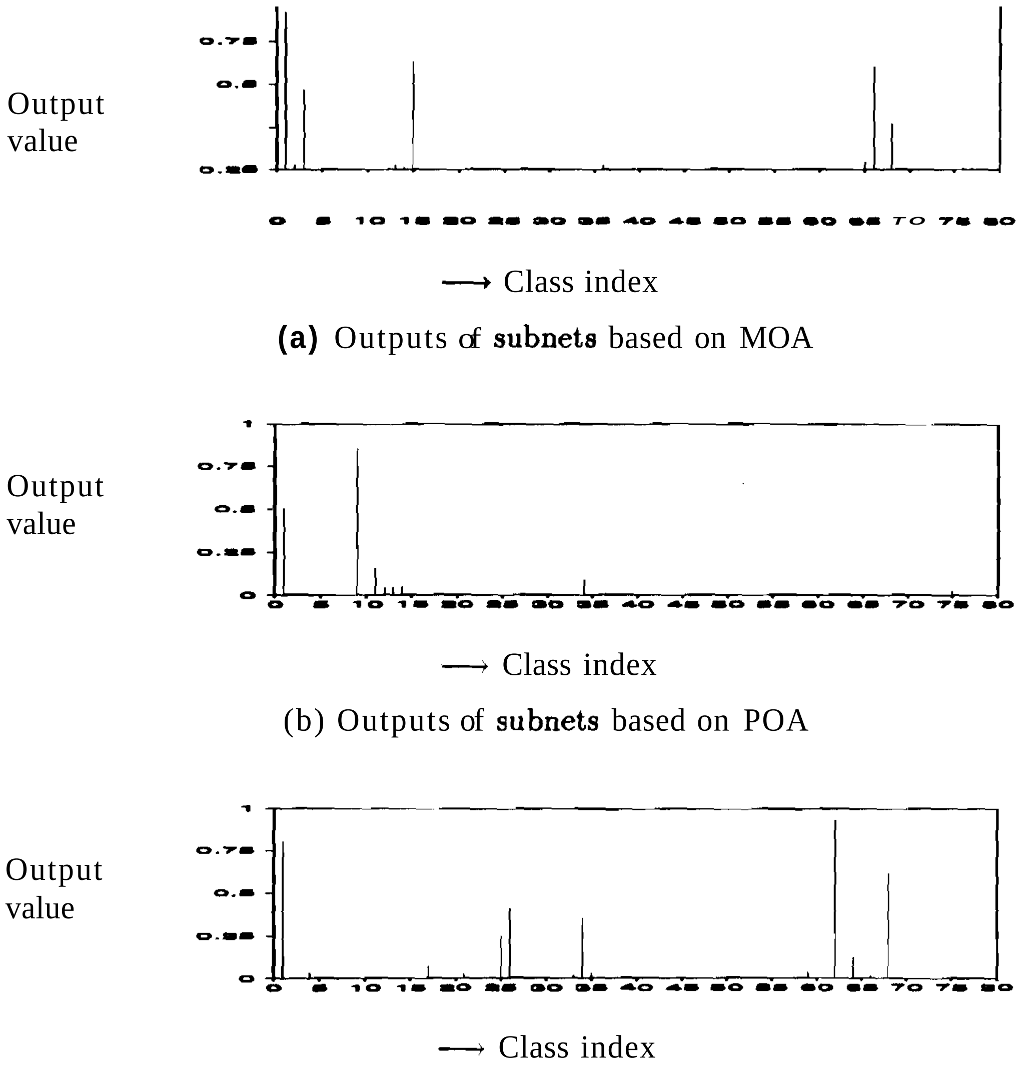
SCV	Grouping Criterion			SCV	Grouping Criterion		
Class	**MOA**	**POA**	Vowel	Class	MOA	**POA**	Vowel
ka	**33**	**16**	**16**	**ṭa**	50	50	58
kha	**66**	58	**33**	**ṭha**	58	50	**33**
ga	**16**	25	50	da	**33**	58	**16**
gha	0	**16**	**0**	dha	**41**	0	25
ki	8	8	**0**	**ṭi**	25	8	**16**
khi	25	25	**33**	**ṭhi**	**33**	58	50
gi	25	8	**41**	di	25	**33**	**16**
ghi	**16**	**16**	25	dhi	**41**	**33**	0
ku	25	25	**41**	**ṭu**	58	100	75
khu	**41**	**33**	**33**	**ṭhu**	25	**66**	50
gu	25	25	**0**	du	**33**	50	58
ghu	8	0	**33**	dhu	25	33	**16**
ke	50	50	**41**	**ṭe**	**33**	**41**	**66**
khe	**16**	0	8	the	0	58	25
ge	25	**16**	**0**	de	8	**41**	8
ghe	**33**	25	8	dhe	58	58	25
ko	**33**	50	**41**	**ṭo**	**41**	75	**16**
kho	25	**33**	**41**	tho	**41**	**83**	**33**
go	25	**16**	25	do	**16**	58	**41**
gho	25	25	**8**	dho	**33**	58	50
ta	**41**	**83**	50	pa	**33**	58	58
tha	**16**	**16**	**0**	pha	**33**	**41**	**41**
da	**41**	**16**	**16**	ba	25	**16**	**66**
dha	**16**	25	**16**	bha	**33**	**33**	8
ti	**41**	**41**	**41**	pi	8	0	25
thi	50	**41**	50	phi	**66**	**66**	**66**
di	25	0	0	bi	**16**	**16**	8
dhi	0	25	8	bhi	25	25	**33**
tu	**66**	50	25	pu	**16**	**33**	58
thu	**33**	25	8	phu	**33**	**33**	25
du	**16**	8	**33**	bu	58	50	**66**
dhu	**16**	**33**	50	bhu	25	**0**	**66**
te	50	**41**	**33**	pe	8	**66**	25
the	**33**	**41**	**41**	phe	25	**66**	50
de	25	**41**	8	be	25	8	**16**
dhe	8	25	**41**	bhe	**41**	25	**33**
to	8	25	25	**po**	25	58	0
tho	**33**	8	**16**	pho	8	50	50
do	8	**41**	**16**	bo	**16**	**16**	**16**
dho	25	25	8	bho	**41**	**33**	33

Table 4.12: Percentage of the total number of test patterns of all the SCV classes that have been correctly classified by different number of modular networks. Only a small percentage of the patterns have been correctly classified by all the three or even two of the three networks. This behaviour is observed for all the three durations.

Segment Duration	All Three Networks	Only Two of the Networks	Only One of the Networks	None of the Networks
200ms	8.7	18.2	31.7	41.4
150ms	4.7	15.6	29.0	50.7
100ms	2.4	10.7	23.8	63.1

The above results clearly show that only a small percentage of the total number of test patterns have been correctly classified by all the three modular networks. It has been observed that even though all the three networks do not give the largest output value for the class of a given pattern, all of them give a significantly large value for that class. This is illustrated in Fig.4.5, where the outputs of subnets based on the three grouping criteria for an input utterance of /ka/ are shown. The classes corresponding to the indices used in this figure are given in Table-4.13.

It can be seen from Fig.4.5 that only the modular network based on MOA grouping correctly classifies the input utterance because the output for the class /ka/ (with the class index of 1) is the largest among the outputs for all 80 classes. The networks based on POA and vowel groupings do not classify correctly. Even though the outputs of subnets in these two networks are not the largest for the class /ka/, they are significantly large. It is also interesting to note that though the POA grouping gives the largest output value for the class /ta/ (with the class index of 9), the outputs of subnets based on MOA and vowel groupings for that class are insignificant. These observations suggest that it is possible to improve the performance by properly combining the evidences available in the outputs of subnets based on different grouping



(a) Outputs of subnets based on MOA

(b) Outputs of subnets based on POA

(c) Outputs of subnets based on Vowel

Figure 4.5: Outputs of subnets based on different grouping criteria for an input utterance belonging to the SCV class /ka/. The classes corresponding to the indices are given in Table-4.13. The output value for the class /ka/ (with the index of 1) is large for all the threegrouping criteria.

Table 4.13: The classes corresponding to the indices used in Fig.4.5.

Index	Class	Index	Class	Index	Class	Index	Class	Index	Class
1	ka	17	ki	33	ku	49	ke	65	ko
2	kha	18	khi	34	khu	50	khe	66	kho
3	ga	19	gi	35	gu	51	ge	67	go
4	gha	20	ghi	36	ghu	52	ghe	68	gho
5	ṭa	21	ṭi	37	ṭu	53	ṭe	69	ṭo
6	ṭha	22	ṭhi	38	ṭhu	54	ṭhe	70	ṭho
7	ḍa	23	ḍi	39	ḍu	55	ḍe	71	ḍo
8	ḍha	24	ḍhi	40	ḍhu	56	ḍhe	72	ḍho
9	ta	25	ti	41	tu	57	te	73	to
10	tha	26	thi	42	thu	58	the	74	tho
11	da	27	di	43	du	59	de	75	do
12	dha	28	dhi	44	dhu	60	dhe	76	dho
13	pa	29	pi	45	pu	61	pe	77	po
14	pha	30	phi	46	phu	62	phe	78	pho
15	ba	31	bi	47	bu	63	be	79	bo
16	bha	32	bhi	48	bhu	64	bhe	80	bho

criteria. In the next chapter, we propose a constraint satisfaction neural network model to combine evidences in the outputs of the all the subnets.

4.7 Summary and Conclusions

In this chapter, we have addressed the issues in developing classifiers for all the SCV classes in Indian languages. We have considered different criteria for grouping the classes and developed modular neural networks for classification. Isolated utterance data has been used for training and testing. Different segment durations have been considered for deriving the fixed duration patterns from isolated utterances. The major results of the studies carried out in this chapter are summarized in Table-4.14.

The performance of subnets based on differentgrouping criteria has been evaluated for different classification models. The subnets using multilayer perceptron model gave a better performance compared to the time delay neural network and hidden

Table **4.14:** Summary of the major results of the studies on development of modular networks for all the SCV classes.

1. Different criteria based on the phonetic description of the SCV classes have been considered for grouping the 80 SCV classes of Hindi into subgroups.

2. Modular networks have been developed for different grouping criteria using MLP, TDNN and DHMM models.

3. The classification performance of subnets based on different grouping criteria and using different models has been compared. The subnets based on POA grouping gave the best average performance. The subnets using MLP models gave a better performance compared TDNN and DHMM models.

4. The analysis of the performance of subnets has shown that about 50% of the total number of errors is due to misclassification of the place of articulation of stop consonants.

5. The analysis of the performance of subnets on test data sets of each SCV class has shown that subnets based on different grouping criteria gave a different performance for many SCV classes.

6. The performance of subnets has been compared for different segment durations used in deriving the patterns from isolated utterances. The performance for the duration of **200** ms is better than the performance for the durations of **150** ms and **100** ms.

7. The modular networks in classification of all the SCV classes gave a significantly poorer performance compared to the performance of the subnets. This is mainly due to the large number of classes and also due to the simple method used for processing the outputs of the subnets. The performance of the modular networks reflects the performance of the subnets used in them.

8. The modular networks based on different grouping criteria gave different performances on test data sets of each of SCV class. This indicates the need for combining the evidences available in the outputs of all the subnets.

Markov models. The performance of subnets based POA grouping gave a better performance compared to the other two grouping criteria. The duration of 200 ms used for deriving the fixed duration patterns from isolated utterances has given a better performance than the durations of 150 ms and 100 ms. The analysis of the results has shown that the performance for the aspirated SCV classes is significantly less than the performance for the unaspirated SCV classes. It is necessary to evolve better techniques for deriving patterns from aspirated SCV utterances. The modular neural network architecture allows one to use suitable preprocessing methods for different subgroups to derive the patterns and improve the performance.

It is observed that the performance of the modular networks is significantly less than the average performance of the subnets. This is mainly because of the simple method used to process the outputs of the subnets. In the next chapter, we propose a constraint satisfaction model in which the knowledge about similarities among the SCV classes is represented in the form of constraints and they are used to combine the evidences in the outputs of the subnets.

Chapter 5

CONSTRAINT SATISFACTION MODEL FOR CLASSIFICATION OF SCV UTTERANCES

5.1 Introduction

In the previous chapter, an architecture based on modular neural networks has been developed for large number of classes. In that architecture the 80 SCV classes are divided into subgroups and a separate neural network (subnet) is trained for each subgroup. For classification of a given SCV utterance, the pattern derived from it is input to all the subnets, and the outputs of the subnets are processed to assign the class of the largest output value to it. Though the classification performance of the subnets for the classes in their subgroups is high (about 65 to 75%), the performance of the modular networks is significantly low (about 30 to 35%). The main reason is that the outputs of the subnets are combined by simply choosing the class with the largest output value. Because of the similarities among several classes, many classes in each subgroup are close to one another. In addition, each subnet is not trained to discriminate all the classes. Confusability among the classes can be resolved to some extent by using the knowledge of the classes. This knowledge can be incorporated as constraints to be met by these classes. A constraint satisfaction model [80] that tries to satisfy as many of these constraints as possible can be used to process the outputs of the subnets. The advantage is that it will work even if some of the constraints derived from acoustic-phonetic knowledge are weak, conflicting and erroneous.

In this chapter we propose a feedback neural network that contains a node for each of the 80 SCV classes. The weight for the connection between a pair of nodes in the network is determined based on the similarity between the classes of the nodes. The similarity between two SCV classes is determined from the knowledge about the differences in their speech production features and also from the confusability between them indicated by the performance of the subnets.

The studies presented in the previous chapter have shown that the subnets based on different grouping criteria give different performance for an SCV class. In the present chapter, we propose a constraint satisfaction model consisting of a feedback network for each of the three grouping criteria. The model combines the multiple evidences available from the subnets of the three criteria for an SCV utterance to decide its class. In this model, the multilayer perceptron network for each subnet is interpreted as a set of nonlinear filters tailored to its subgroup. The output of the filters for an utterance is viewed as a feature vector representing the utterance. The distribution of the feature vectors for a class may be different for each grouping criterion. The feature vector of an utterance is given as input to the feedback network corresponding to that group. Then the constraint satisfaction model is allowed to relax to an equilibrium state. The resulting state represents a situation where the constraints are satisfied to maximum extent for the given input to the feedback networks. A stable state is expected to be close to the correct one even though the constraints are weak due to partial knowledge used in deriving the constraints, and also even if the representation of the discriminatory information in the feature vectors is poor.

The organization of this chapter is as follows: The next section explains the interpretation of multilayer perceptrons as nonlinear feature extractors. Section 5.3

describes the method used for deriving the weights for the connections in the feedback networks. The proposed constraint satisfaction model is described in Section 5.4. The operation of the constraint satisfaction model, the relaxation strategy and the interpretation of the stable state of the network are also discussed in this section. The classification of SCV utterances by the constraint satisfaction model is presented in Section 5.5.

5.2 Neural Networks as Nonlinear Feature Extractors

A multilayer perceptron is a Multilayer Feedforward Neural Network (MLFFNN). The MLFFNN trained for a subgroup of classes is considered as a filter designed in such a way that it provides discrimination among the classes. One such network is used for each subgroup consisting of about 16 or 20 SCV classes depending on the grouping criterion used. Thus there are 16 or 20 filters in each subgroup and **80** filters for each grouping criterion. It may be noted that each SCV class occurs with a different subgroup for each of the three groupings. We can also interpret the network as a filter set tailored to the classes in a subgroup. This is like Gabor filters used for texture classification where the filters are tailored to the characteristics of the texture classes under consideration [81]. The characteristics to be optimized in the case of Gabor filters are resolution, orientation and spatial frequency.

The shape of the decision surface formed for a class by an MLFFNN will vary depending on the other classes in the subgroup. This behaviour is illustrated for an arbitrary 2-dimensional pattern space in Fig.5.1. In this figure the regions for **10** classes are shown. We consider one class for which there are eight classes that are close to it. The region for the class under consideration is shown in dark shade. When this class is grouped with three other classes that are close to it, the decision surface formed for the class is dependent on which three classes are present in the subgroup.

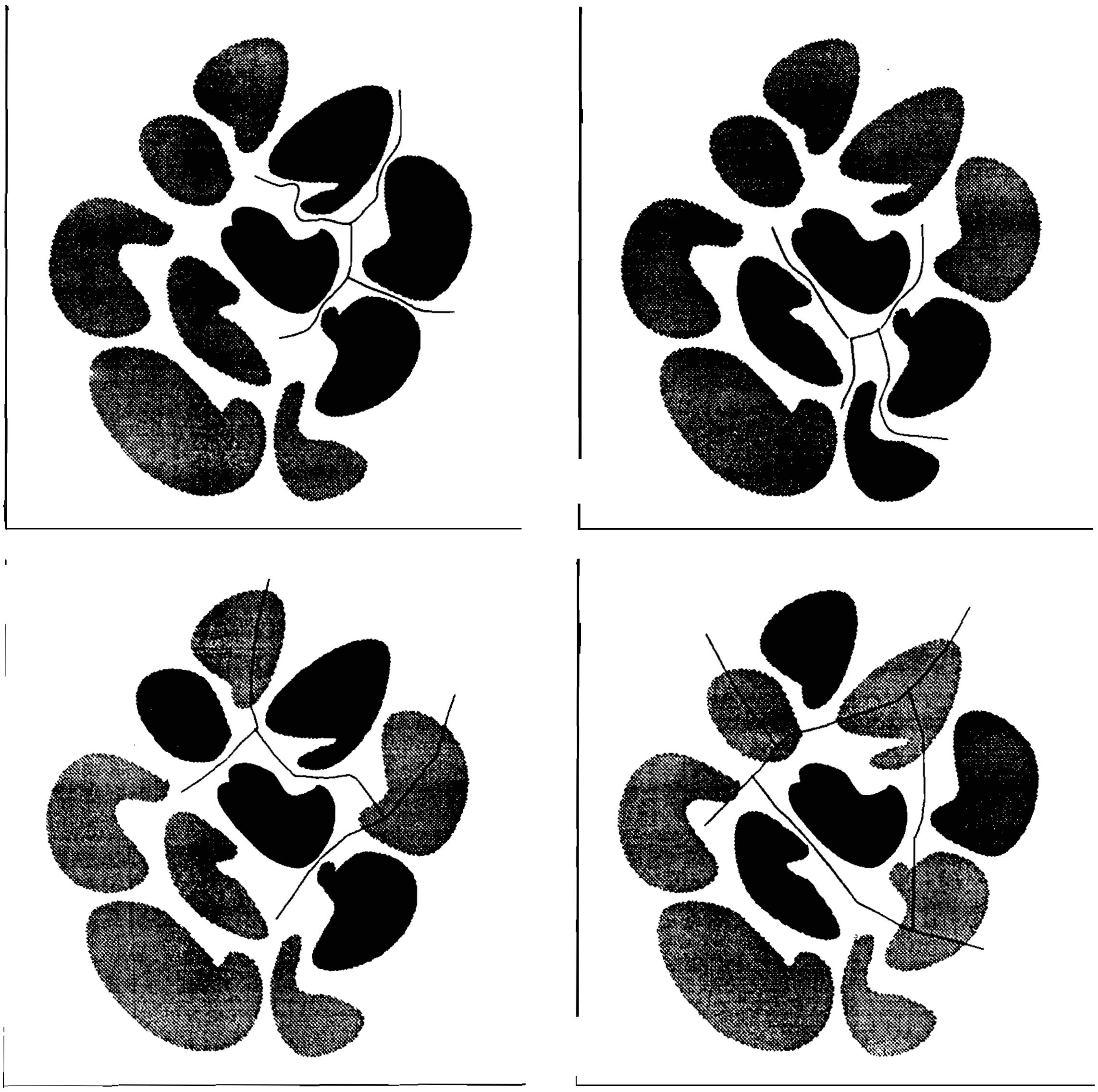


Figure 5.1: Illustration of the effect of grouping a class with different subsets of classes on the decision surfaces formed. An arbitrary 2-dimensional pattern space is used to explain the effect of different ways of grouping. The region of a class under consideration is shown in dark shade. The regions of the classes with which the class under consideration is grouped are shown in a different shade. Typical decision surfaces expected to be formed around the region of the class under consideration are shown for 4 different subsets of classes with which the class is grouped.

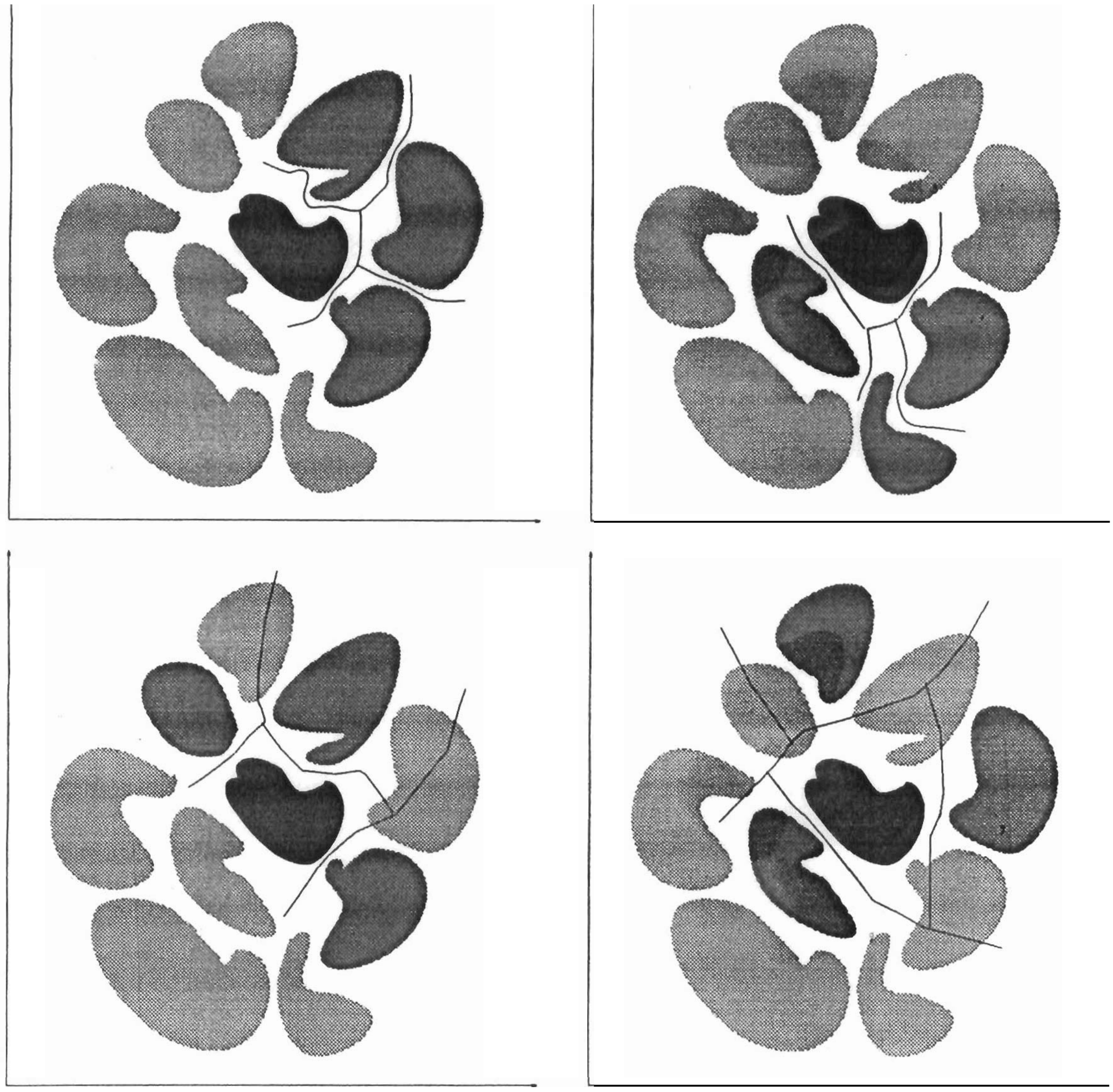


Figure 5.1: Illustration of the effect of grouping a class with different subsets of classes on the decision surfaces formed. An arbitrary 2-dimensional pattern space is used to explain the effect of different ways of grouping. The region of a class under consideration is shown in dark shade. The regions of the classes with which the class under consideration is grouped are shown in a different shade. Typical decision surfaces expected to be formed around the region of the class under consideration are shown for **4** different subsets of classes with which the class is grouped.

Typical decision surfaces that are expected to be formed for MLFFNNs trained for four different subgroups containing the class are shown in the figure. The shapes of the decision surfaces for the class are different in each of the MLFFNNs. Therefore the different MLFFNNs are likely to give different output values for the class of the input pattern.

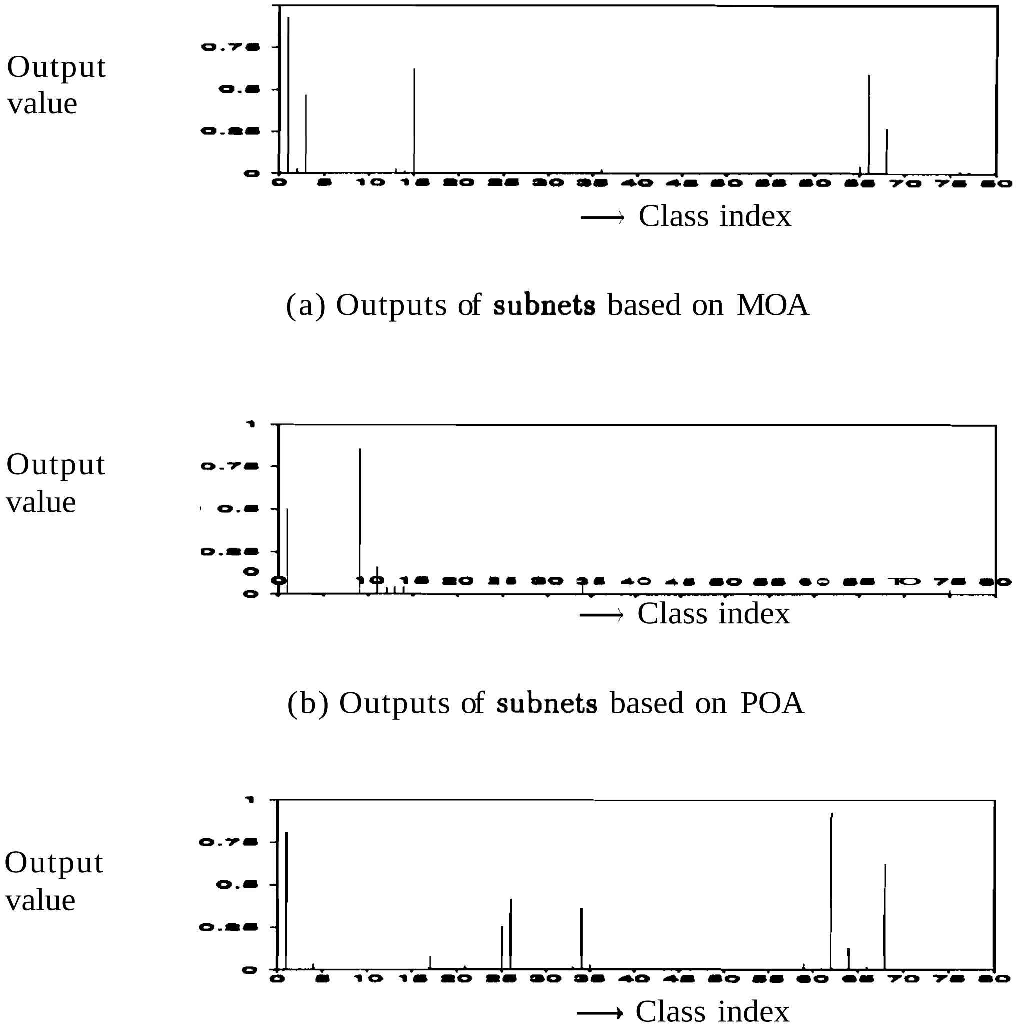
This behaviour has been observed in the outputs of the subnets. The outputs of the subnets for an input utterance of the class /ka/ are shown in Fig.5.2. It can be seen that the output value for the class /ka/ (with the class index of 1) is different for the three subnets.

Normally the trained MLFFNNs are used directly as classifiers for subgroups of classes. But the concept of filter interpretation provides greater flexibility and robustness in the development of a classifier for all the SCV classes. Once the MLFFNNs are trained, then they are used as nonlinear filters. The outputs of the filters for each subgroup for a given training sample is considered as a feature vector. The distribution of the feature vectors is obtained for each class from a second training set data. The distribution is represented in terms of a mean vector and a variance parameter derived from the feature vectors for the class.

The outputs of the sets of filters designed in this section are input to the feedback networks in the constraint satisfaction model. The next section will describe the feedback networks and explain the method of determining the weights for the connections in the networks. These weights represent the constraints, and they are derived using the acoustic-phonetic knowledge and the performance statistics of the subnets.

5.3 Feedback Networks for Different Grouping Criteria

We first build three different feedback networks, one for each of the three grouping criteria. Since the SCV classes within a subgroup have been designed to compete



(a) Outputs of **subnets** based on MOA

(b) Outputs of **subnets** based on POA

(c) Outputs of **subnets** based on Vowel

Figure 5.2: Difference in the outputs of **subnets** based on different. grouping criteria for an input utterance belonging to the SCV class /ka/. It can be noted that the output values for the class /ka/ (with the class index of 1) are different for each grouping criterion. This is mainly due to different constitution of the subgroups containing the class /ka/ for different grouping criteria.

amongst themselves during training of the MLFFNN for that subgroup, we provide excitatory connections between the nodes corresponding to the classes in a subgroup. All the connections across the subgroups are made inhibitory. The weights for the excitatory and inhibitory connections have been derived from the confusion matrices obtained from the classification performance of subnets.

We first obtain the confusion matrices for different manners of articulation, different places of articulation and for different vowels. The confusion matrix for different manners of articulation (MOAs) of stop consonants in SCVs is obtained as follows. We determine the percentage of the total number of patterns belonging to the classes with a particular MOA, say Unvoiced-Unaspirated (UVUA), that were classified as belonging to the classes in different manner subgroups, namely, UVUA, Unvoiced-Aspirated (UVA), Voiced-Unaspirated (VUA) and Voiced-Aspirated (VA). The confusion matrix for different manners of articulation is given in Table-5.1(a) below. The confusion matrix for different places of articulation is given in Table-5.1(b). The confusion matrix for different vowels in SCVs is given in Table-5.1(c).

The confusion matrices are used to derive symmetric similarity matrices. The similarity matrix for different manners of articulation is obtained as follows. From Table-5.1(a), it is noted that 2.5% of the total number of patterns belonging to the SCV classes with UVUA as the MOA are classified as belonging to the classes with UVA as the MOA, and 3.1% of the total number of patterns belonging to the the classes with UVA are classified as belonging to the classes with UVUA as the MOA. On an average, 2.8% of the total number of patterns belonging to the classes with UVUA and UVA as MOA are misclassified due to the confusion between these two manners. The average percentage is used to determine the similarity measure in the range 0.0 to 1.0. The similarity between UVUA and UVA is indicated as 2.8/100 which is rounded to 0.03. The similarity matrix for different manners of articulation

Table **5.1:** Confusion matrices for (a) different manners of articulation of stop consonants in SCVs, (b) different places of articulation of stop consonants in SCVs and (c) different vowels in SCVs.

(a) Confusion matrix for different manners of articulation.

MOA	UVUA	UVA	VUA	VA
UVUA	**86.9**	**2.5**	**7.7**	2.9
UVA	**3.1**	**84.2**	**4.4**	**8.3**
VUA	**4.2**	**3.3**	**78.3**	**14.2**
VA	**0.4**	**7.9**	**9.6**	**82.1**

(b) Confusion matrix for different places of articulation.

POA	Velar	Alveolar	Dental	Bilabial
Velar	**72.1**	**9.0**	**7.1**	**11.8**
Alveolar	**6.0**	**75.4**	**7.3**	**11.3**
Dental	**8.1**	**12.9**	**67.7**	**11.3**
Bilabial	**4.2**	**7.5**	**8.5**	**79.8**

(c) Confusion matrix for different vowels.

Vowel	/a/	/i/	/u/	/e/	/o/
/a/	89.3	1.0	2.9	0.8	6.0
/i/	0.5	85.7	3.1	9.9	0.8
/u/	1.3	1.6	82.3	1.3	13.5
/e/	1.3	6.8	1.3	90.3	0.3
/o/	4.4	0.0	18.8	0.0	76.8

is given in Table-5.2(a), for different places of articulation in Table-5.2(b) and for different vowels in SCVs in Table-5.2(c).

The similarity measures are used to determine the weights for the excitatory and inhibitory connections in the feedback networks. An excitatory connection is provided between nodes of two SCV classes within a subgroup if they differ in only MOA or POA or vowel characteristic. The weight of an excitatory connection is equal to the sirnilarity measure between the differing production features of the two classes. For example, in grouping based on MOA, the SCV class /ka/ belongs to UVUA subgroup. Of the 20 SCV classes present in this subgroup (/ka/, /ṭa/, /ta/, /pa/, /ki/, /ṭi/,

Table 5.2: Similarity matrices for (a) different manners of articulation of stop consonants in SCVs, (b) different places of articulation of stop consonants in SCVs and (c) different vowels in SCVs.

(a) Similarity matrix showing the closeness between different manners of articulation.

MOA	UVUA	UVA	VUA	VA
UVUA	0.87	0.03	0.06	0.02
UVA	0.03	0.84	0.04	0.08
VUA	0.06	0.04	0.78	0.12
VA	0.02	0.08	0.12	0.82

(b) Similarity matrix showing the closeness between different places of articulation.

POA	Velar	Alveolar	Dental	Bilabial
Velar	0.72	0.08	0.08	0.08
Alveolar	0.08	0.75	0.10	0.09
Dental	0.08	0.10	0.68	0.10
Bilabial	0.08	0.09	0.10	0.80

(c) Similarity matrix showing the closeness between different vowels.

Vowel	/a/	/i/	/u/	/e/	/o/
/a/	0.89	0.01	0.02	0.01	0.05
/i/	0.01	0.86	0.02	0.08	0.00
/u/	0.02	0.02	0.82	0.01	0.16
/e/	0.01	0.08	0.01	0.90	0.00
/o/	0.05	0.00	0.16	0.00	0.77

/ti/, /pi/, /ku/, /ṭu/, /tu/, /pu/, /ke/, /ṭe/, /te/, /pe/, /ko/, /ṭo/, /to/ and /po/), an excitatory connection is provided between /ka/ and each of the following seven classes only : /ṭa/, /ta/, /pa/, /ki/, /ku/, /ke/ and /ko/. The remaining 12 classes in this subgroup differ with /ka/ in both POA and vowel, and hence no connection is provided between the nodes of /ka/ and these 12 classes. The weight for the excitatory connection between /ka/ and /ki/ is 0.01 which is the similarity measure between vowels /a/ and /i/ as given in Table-5.2(c).

An inhibitory connection is provided between classes in different subgroups only if the two classes differ either in MOA or POA or vowel only. For the earlier example of class /ka/ in the grouping based on MOA, an inhibitory connection is provided between /ka/ in UVUA subgroup and each of the following classes: /kha/ in UVA, /ga/ in VUA and /gha/ in VA subgroup. All the other classes in UVA, VUA and VA subgroups differ with /ka/ not only in MOA but also in POA or/and vowel. The weight for an inhibitory connection is inversely proportional to the similarity measure between the differing production features of the two classes. If the similarity measure is C (in the range 0.0 to 1.0), then the inhibitory weight W is assigned as follows:

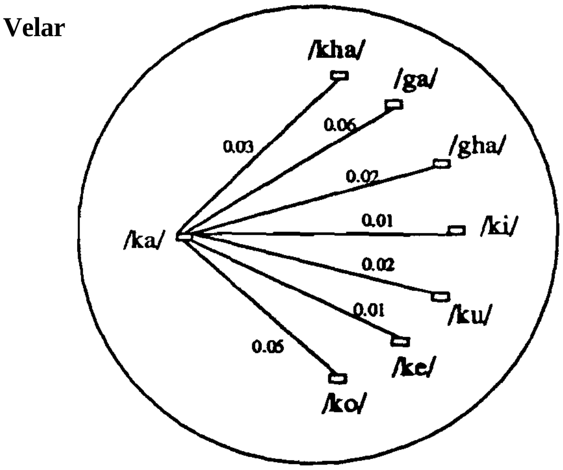
$$W = -\frac{1}{100 * C} \tag{5.1}$$

If the closeness measure C is less than 0.01, then the corresponding inhibitory weight is assigned as -1.0. The weights of the connections for the class /ka/ in the feedback networks for different grouping criteria are given in Table-5.3.

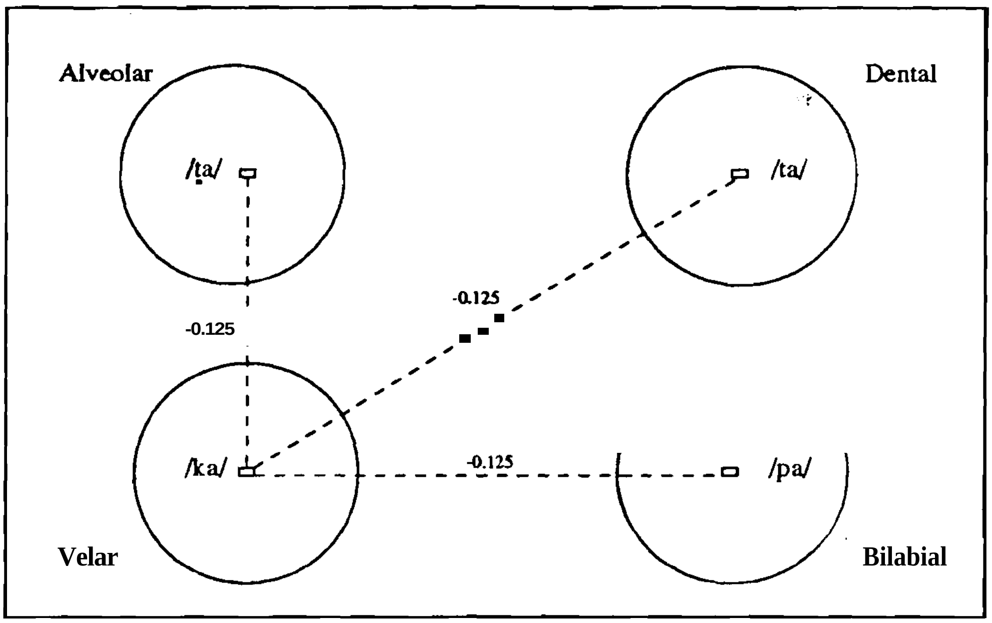
The connections in the feedback network for the grouping criterion of POA are illustrated in Fig.5.3. The excitatory connections for the class /ka/ in the 'Velar' subgroup are shown in Fig.5.3(a) and the inhibitory connections for the class are shown in Fig.5.3(b).

The main function of each feedback network is to enhance the evidence available from the filters for the class of the input utterance by giving positive contributions from evidences for the classes close to it in a subgroup, and to reduce the evidence for the classes which are in the other subgroups but are close to it. The weights of the connections based on similarities among classes help the feedback network to perform its function.

Each unit in a feedback network is associated with a mean vector μ and a variance parameter σ representing the distribution of feature vectors for the class of the unit. The mean vector and the variance parameter are obtained from a second training set data. A training pattern belonging to the class of the unit is input to the subnet for



(a) Excitatory connections for the class /ka/ in the POA feedback network



(b) Inhibitory connections for the class /ka/ in the POA feedback network

Figure 5.3: Connections for the class /ka/ in the POA feedback network. The excitatory connections for the class /ka/ in the 'Velar' subgroup are shown in (a). The inhibitory connections for the class /ka/ are shown in (b).

Table 5.3: Illustration of weights of connections for class /ka/ in the feedback networks for different grouping criteria.

Grouping Criterion	Excitatory Connections		Inhibitory Connections	
	Class	Weight	Class	Weight
MOA	/ṭa/	0.08	/kha/	-0.33
	/ta/	0.08	/ga/	-0.16
	/pa/	0.08	/gha/	-0.50
	/ki/	0.01		
	/ku/	0.02		
	/ke/	0.01		
	/ko/	0.05		
POA	/kha/	0.03	/ṭa/	-0.125
	/ga/	0.06	/ta/	-0.125
	/gha/	0.02	/pa/	-0.125
	/ki/	0.01		
	/ku/	0.02		
	/ke/	0.01		
	/ko/	0.05		
Vowel	/ṭa/	0.08	/ki/	-1.0
	/ta/	0.08	/ku/	-0.5
	/pa/	0.08	/ke/	-1.0
	/kha/	0.03	/ko/	-0.2
	/ga/	0.06		
	/gha/	0.02		

the subgroup containing the class. The output of the subnet is used to form a feature vector. The dimension of the feature vector is same as the number of classes in the subgroup. If $\mathbf{y}_i$ is the feature vector obtained for the ith training pattern and N is the number of training patterns for each class, then the kth element of the mean vector, μ_k, is computed as follows:

$$\mu_k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_{ik} \tag{5.2}$$

where $\mathbf{y}_{ik}$ is the kth element of y;. The variance parameter a is computed from the mean vector and the feature vectors for the N training patterns as follows:

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}(y_{ik} - \mu_k)^2 \qquad (5.3)$$

where M is the dimension of the feature vectors and the mean vector.

The mean vector and the variance parameter are computed for each of the 80 SCV classes and for each of the three grouping criteria during the second level of training: For classification of an SCV utterance, the pattern belonging to the utterance is input to all the subnets. The outputs of the subnets are input to the feedback network corresponding to their grouping.

5.4 Constraint Satisfaction Model for Classification of SCVs

The feedback networks for different grouping criteria interact with each other through a pool of units, called instance pool [82]. There are as many (80) units in the instance pool as the number of SCV classes. Each unit in the instance pool (for example, the unit, corresponding to class /ka/) has a bidirectional excitatory connection with the corresponding units in the feedback networks (for example, units corresponding to /ka/ in MOA group, /ka/ in POA group and /ka/ in Vowel group). Units within the instance pool compete with one another and hence are connected by a fixed negative weight (-0.2). Thus the three feedback networks along with the instance pool constitute the constraint satisfaction model reflecting the known speech production knowledge of the SCVs as well as the knowledge derived in the organization of the classes for training the subnets. The constraint satisfaction model developed for classification of SCVs is shown in Fig.5.4.

The operation of the constraint satisfaction model is as follows: Each unit j in the constraint satisfaction model computes the weighted sum of inputs from the other units (net,) in the model. An external input for each of the units in the feedback

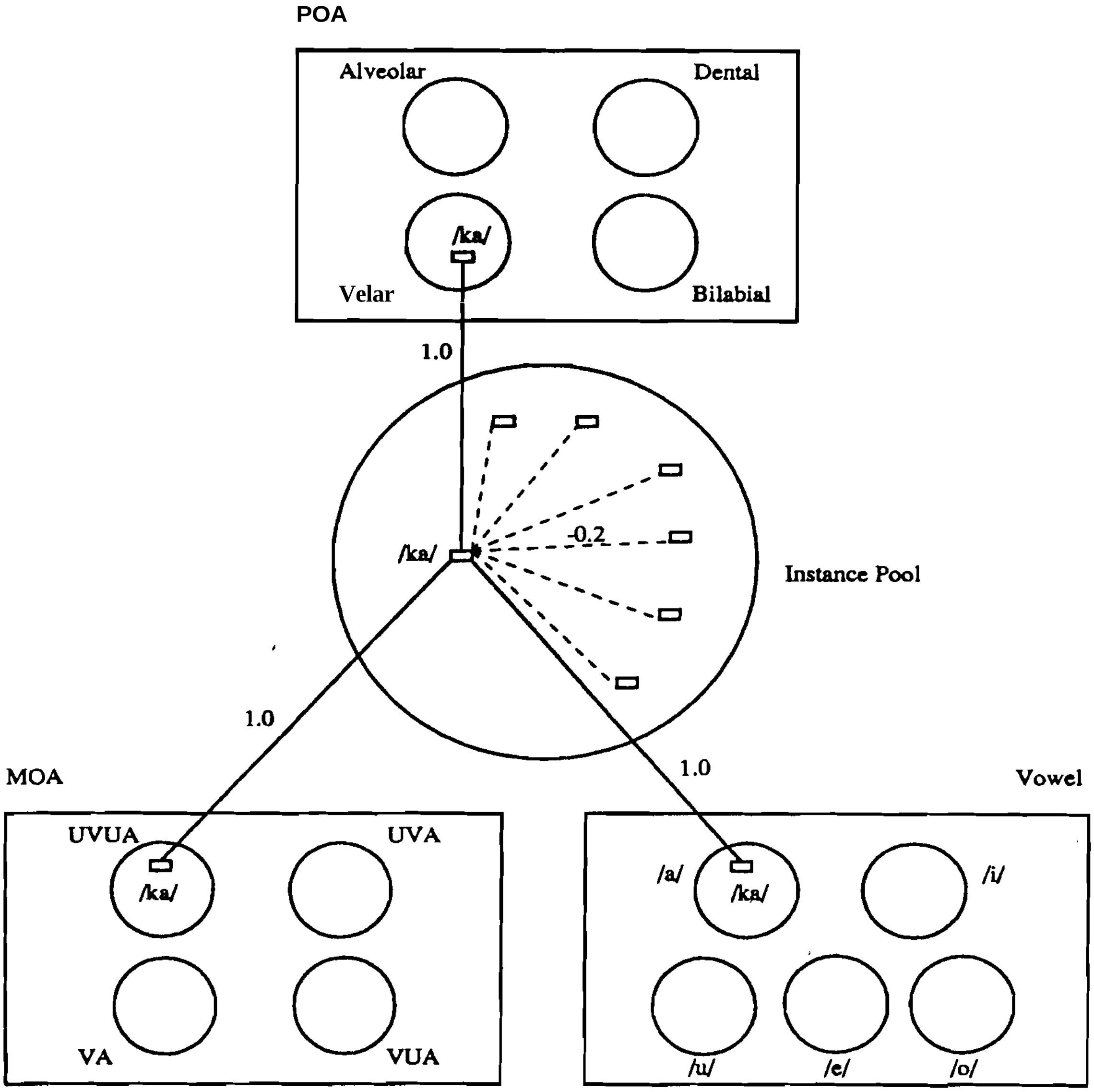


Figure 5.4: Constraint satisfaction model for classification of SCV utterances. The constraint satisfaction model consists of three feedback networks for three grouping criteria, and an instance pool through which the feedback networks interact. The instance pool has one node for each class. The instance pool node for a class is connected to the nodes of that class in the three feedback networks by a bidirectional excitatory connection. A node in the instance pool has a bidirectional inhibitory connection with all the other nodes in the pool.

networks is provided as bias derived from the 16- or 20-dimensional feature vector of the subgroup to which the unit belongs to. For a given input pattern, the output of an MLFFNN is considered as a feature vector denoted by x. Each unit j is associated with a mean vector μ_j and a variance parameter σ_j. Then the external bias for the unit, $bias_j$, is computed as follows:

$$bias_j = \frac{1}{\sqrt{(2\pi)^M {\sigma_j}^2}} e^{-\frac{distance}{2}} \tag{5.4}$$

$$\text{distance} = \frac{|\mathbf{x} - \mu_\mathbf{j}|^2}{{\sigma_j}^2} \tag{5.5}$$

$$|\mathbf{x} - \mu_\mathbf{j}|^2 = \frac{1}{M} \sum_{i=1}^{M} (\mathrm{x_i} - \mu_{\mathrm{ji}})^2 \tag{5.6}$$

where M is the dimension of the feature and mean vectors, $\mathbf{x_i}$ and $\mu_{\mathbf{ji}}$ are the ith elements of the feature vector x and the mean vector $\mu_\mathbf{j}$ respectively.

The net input to the unit j is given by

$$netinput_j = \alpha * bias_j + \beta * net_j + \theta \tag{5.7}$$

where a, β and θ are constants in the range 0.0 to 1.0, chosen empirically by trial and error. The output function for the units is a sigmoid function and is given by

$$output_j = \frac{1}{1 + e^{-k * netinput_j}} \tag{5.8}$$

where k is a constant that determines the slope of the sigmoid curve.

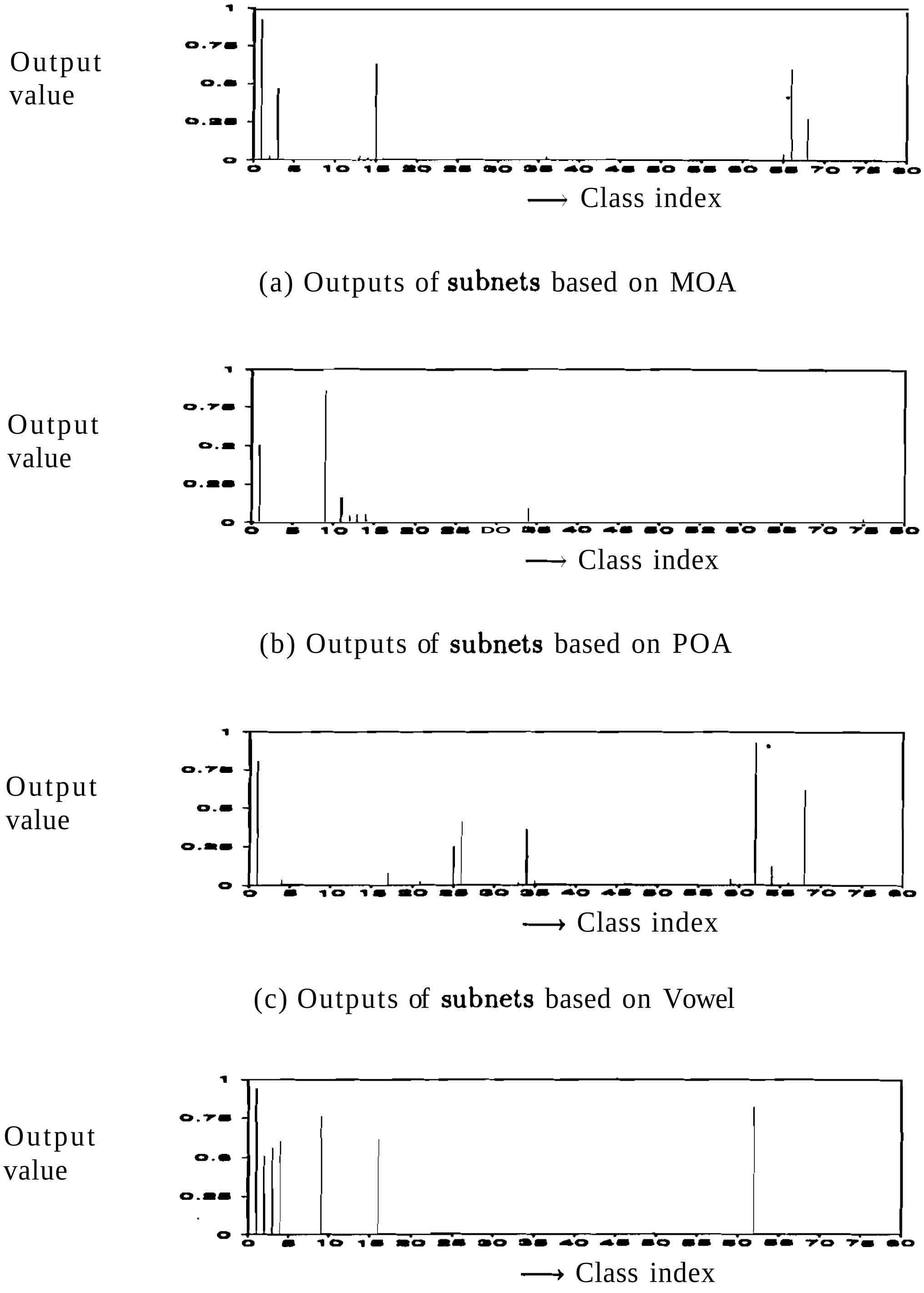
The constraint satisfaction model is initialized as follows: When a new pattern is presented to the MLFFNNs, the feature vectors, x's, for all the MLFFNNs are obtained. The outputs of the units in the feedback networks for whom the corresponding feature vector element value is above a threshold 6 (0.3) are initialized to +1.0 and the outputs of all other units in the feedback networks are initialized to 0.0. The bias for a unit in the instance pool is computed from the net input to the unit after the feedback networks are initialized. The output of a unit in the instance pool is initialized to +1.0, if the net input to the unit is greater than 0.0. The constraint satisfaction model is then allowed to relax until a stable state is reached for a given input. Deterministic relaxation method is used. In this method a unit in the model is chosen at random and its output is computed. This method is continued until there is no significant change in the outputs of the units. At a stable state of the model, the outputs of the instance pool units are interpreted to determine the class of the input pattern.

The class of the instance pool unit with the largest output value is assigned as the class of the input utterance. Because of similarity amongst SCV classes, we consider the cases in which the correct class can be among the classes corresponding to the K largest output values. In the next section, we present the classification performance of the constraint satisfaction model for Case-1, Case_2, Case-3, and Case-4, corresponding to K = 1, 2, 3, and 4, respectively.

If the feature vectors for an input pattern are considered as evidences for the classes available from subnets based on different grouping criteria, then the outputs of the instance pool units in the final state of the model can be considered as the combined evidence for each class after satisfaction of as many constraints as possible. The feature vectors for an input pattern belonging to the class /ka/ are plotted in Fig.5.5. The outputs of the instance pool units in the final state of the model are also shown in the figure.

5.5 Results and Discussion

In the previous chapter, we have considered different durations of segments for deriving the patterns from isolated utterances of SCVs, and compared the classification performance of the subnets and the modular networks. In this section we present



(a) Outputs of subnets based on MOA

(b) Outputs of subnets based on POA

(c) Outputs of subnets based on Vowel

(d) Outputs of the instance pool units

Figure 5.5: Outputs of subnets based on different grouping criteria and the instance pool units in the constraint satisfaction model (CSM) for an input utterance belonging to the SCV class /ka/. Multiple evidences available from the three criteria are combined by the CSM using the constraints based on similarities among the classes.

the performance of the constraint satisfaction model for classification of all the SCV classes. The average performance of the constraint satisfaction model on. the test data is given in Table-5.4 for different segment durations and for different cases of decision criterion. The performance of the modular networks is also given in the table for comparison. Here we consider the performance of the modular networks based on POA grouping as they gave the best classification performance among the three groupings.

Table 5.4: Classification performance of the constraint satisfaction model on the test data of all the SCV classes for different segment durations. The performance of the modular network with POA as the grouping criterion is also given for comparison. The performance is given for four different cases of decision criterion. The Case_M for the constraint satisfaction model refers to the criterion that the class of an input pattern is amongst the M largest output values of the instance pool units. The Case_M for the modular network refers to the case when the M largest values amongst the outputs for all the SCV classes are considered in deciding the class. The constraint satisfaction model gives a significantly better performance than the modular network. This behaviour is observed for different segment durations.

Segment Duration	Constraint Satisfaction Model				Modular Network			
	Case-1	Case_2	Case_3	Case-4	Case-1	Case_2	Case_3	Case-4
200ms	65.6	75.0	80.2	82.6	35.1	56.9	69.5	76.6
150ms	54.3	67.5	73.1	76.7	30.8	49.4	60.1	66.7
100ms	39.0	49.5	56.0	60.8	20.6	36.4	47.0	53.9

It is observed that the classification performance of the constraint satisfaction rnodel (CSM) is distinctly better than the performance of the modular networks. The performance of CSM for Case-1 is as high as 65% for the segment duration of 200 ms indicating that the instance pool unit with the largest output value gives the class of the input utterance correctly for 65% of the total number of test utterances. This result is significant considering that the classification is performed by the CSM by discriminating among 80 SCV classes and that many of these classes are similar.

The performance of the CSM increases to about **82%** for the Case-4 of the decision criterion.

The postprocessor in a modular network processes the outputs of the subnets in that network to decide the class. The postprocessor simply assigns the class of the largest output value without using the similarity information available in the other outputs. The modular networks for different groupings operate independent of each other. The evidences available from different modular networks are not used in the classification. In the CSM, the outputs from **subnets** of each grouping criterion are processed by the feedback network for that grouping. Similarities among classes are represented in the weights of the connections in the feedback network. Evidences available from different groupings are combined by letting the feedback networks interact with one another through the instance pool. Therefore the CSM not only uses the knowledge about the similarities among classes but also combines the evidences from multiple sources in performing the classification. It is observed that even the weak evidences available from different sources are enhanced by the CSM during relaxation. The improved classification performance of the CSM is mainly due to its ability to combine evidences from multiple sources.

The results of the studies in the previous chapter have shown that the classification performance of the modular networks is much less than the average performance of subnets. It is observed that the performance of the CSM approaches the performance of subnets. In Table-5.5 we give the performance of the CSM on the test data sets of the classes in different subgroups. The table also gives the performance of subnets for the same data. It is important to note that the CSM performs classification by discriminating among all the **80** classes where as a subnet performs classification by discriminating only among the 16 or **20** classes in its subgroup. It is significant that the performance of the CSM is close to that of the subnets. The performance of the

CSM and the subnets is given for different segment durations used for analysis.

Table 5.5: Classification performance of the constraint satisfaction model and the subnets on test data of SCV classes in subgroups based on different grouping criteria. The comparison of performance is given for different segment durations. The performance of the constraint satisfaction model on data for a subgroup of classes is close to the performance of the subnet for that subgroup. The constraint satisfaction model performs classification by discriminating among all the 80 SCV classes where as the subnet discriminates only among the classes in its subgroup. This behaviour is observed for different segment durations.

Grouping		200ms		150ms		100ms	
Criterion	Subgroup	CSM	Subnet	CSM	Subnet	CSM	Subnet
MOA	UVUA	76.7	77.1	65.0	64.1	53.4	62.7
	UVA	66.3	70.0	39.6	40.4	22.3	31.0
	VUA	62.1	62.9	69.2	65.8	54.8	64.0
	VA	57.5	56.3	43.3	43.8	25.0	35.0
POA	Velar	55.0	54.6	55.6	60.8	42.1	51.5
	Alveolar	73.3	84.2	53.5	62.5	36.9	54.6
	Dental	66.3	77.9	50.2	68.1	31.9	51.0
	Bilabial	67.9	80.0	57.7	71.5	44.6	56.5
Vowel	/a/	68.2	61.5	63.0	69.8	54.9	64.1
	/i/	66.2	66.6	51.6	63.5	29.7	31.3
	/u/	61.5	68.8	54.2	63.0	37.3	45.8
	/e/	65.6	62.0	52.6	61.5	39.1	41.7
	/o/	66.7	67.7	50.0	69.5	33.4	40.4

It is interesting to note that the performance of the CSM some times exceeds even that of the subgroup. For example, consider the performance for the '/a/' subgroup for the segment duration of 200 ms. The performance of the CSM for this subgroup is 68.2%, whereas the performance its subnet is only 61.5%. This is possible because the CSM uses the evidences available from subnets based on all three grouping criteria. A similar behaviour can be noted for the VUA subgroup for the duration of 150 ms.

The performance of the subnets is important for the performance of the constraint satisfaction model. It can be seen that the poorer performance of the CSM for the durations of 150 ms and 100 ms compared to 200 ms is mainly because of the poor performance of the subnets for the reduced durations. Even a marginal improvement in the performance of the subnets can lead to a significant improvement in the performance of the CSM.

5.6 Summary and Conclusions

A summary of the issues addressed and major results of the studies carried out in this chapter are given in Table-5.6. In this chapter, we have proposed a new approach for developing a model for classification of utterances of all the SCV classes. In this approach we proposed a constraint satisfaction model to represent the known constraints of the problem. Trained multilayer feedforward neural networks are used as nonlinear filters to extract features. A second level of training is used to derive the distribution of the features for each class. Since the constraint satisfaction model satisfies a set of weak constraints in the best possible manner, the results are good in most of the cases.

The ability of the CSM to combine multiple evidences is useful for performing speaker independent classification. The subnets are trained for the data collected from multiple speakers and no speaker adaptation is performed by them. Therefore the outputs of subnets for an utterance from a new speaker can be low and hence the evidences available may be weak. Though the performance of subnets is not speaker independent, the operation of the CSM is speaker independent and hence it is expected that the CSM would show a distinct improvement in speaker independent classification over the modular networks.

The main difficulty in improving the classification performance is in the parametric representation of speech. If the representation does not capture the crucial information from the speech signal, then there is no hope of classifying the input cor-

Table 5.6: Summary of the results of studies on development of a constraint satisfaction model for classification of SCV utterances.

1. Neural networks used for building subnets for subgroups of SCV classes have been interpreted as nonlinear feature extractors. This interpretation is useful in looking at the outputs of a subnet for an input SCV pattern as the outputs of the set of nonlinear filters developed during training of a subnet for the classes in its subgroup.

2. A feedback neural network has been developed for each of the three criteria considered for grouping SCV classes. A second level of training has been carried out to obtain the mean feature vector and variance for each SCV class, which are used in computing the bias value for the node of a class in the feedback network. The weights for the connections in the feedback networks are determined using the similarities among SCV classes derived from experimental results.

3. A constraint satisfaction model has been developed in which the three feedback networks interact with one another through an instance pool. The multiple evidences available from different grouping criteria for an input SCV pattern are combined using a relaxation method to determine its class.

4. The constraint satisfaction model gave a significantly better (about 30% higher) performance compared to the modular networks for classification of utterances of all the 80 SCV classes.

5. The performance of the constraint satisfaction model in the classification of the 80 SCV classes approaches the average performance of the subnets in classification of subgroups of the SCV classes.

rectly using a simple classifier. Parametric representation is also a limiting factor for realizing speaker independent classification of SCV utterances. Our studies demonstrate the power of constraint satisfaction models to enhance even the weak evidence available in the input data. The models developed for the classification of isolated utterances of SCVs can be used for spotting SCV segments in continuous speech. In the next chapter, we propose an approach in which the speech signal around the vowel onset points is scanned by the classification models to spot SCVs.

Chapter 6

SPOTTING SCV SEGMENTS IN CONTINUOUS SPEECH

6.1 Introduction

The focus of the studies in the previous chapters has been on developing models for classification of continuous speech segments belonging to a small set of frequently occurring SCV classes, and on developing models for classification of isolated utterances of all the 80 SCV classes. Models based on One-Class-One-Network (OCON) and All-Class-One-Network (ACON) architectures have been considered for small sets of classes. Modular neural network architecture has been considered for classification large number of classes. In this chapter we address the issues in using these classification models for spotting SCV segments in continuous speech. We propose an approach in which the vowel onset points (VOPs) in continuous speech are detected first, and the classification models for SCVs are then used to scan the speech segments around VOPs for spotting SCVs. Spotting SCVs in continuous speech is useful in developing vocabulary independent continuous speech recognition system.

The organization of this chapter is as follows: The main issues in spotting SCVs are discussed in the next section. In section 6.3 we describe a neural network based method for detection of vowel onset points in continuous speech. In section 6.4 we present the studies on spotting SCVs.

6.2 Issues in Spotting SCVs

Strategies for spotting subword units in continuous speech have been based on training a model for each of the classes to classify only the segments of the continuous speech signal belonging to that class and reject all other segments [42]. The models thus trained are then used to scan the speech signal continuously and hypothesize the presence or absence of the corresponding subword units. The hypotheses from the models for all the classes are processed further to hypothesize the subword units present in a given continuous speech signal. We discuss some issues in adopting this strategy and propose an approach for spotting SCVs.

The commonly used approaches for spotting subword units scan the speech signal continuously. In these approaches patterns extracted from fixed duration segments starting at every 5 or 10 ms are given as input to a classifier to determine their classes. In our approach to spotting SCVs, we propose to identify the Vowel Onset Points (VOPs) in continuous speech and scan a portion of the speech signal around each VOP to determine the class of a segment around the VOP. By restricting the scanning to regions around VOPs, the portions of the speech signal not belonging to CV segments are eliminated from consideration thus leading to a significant reduction in the number of false alarms. The resulting false alarms are mainly due to the errors made by the classifiers.

The performance of any spotting approach is mainly dependent on the capability of the classifiers to correctly classify the segments belonging to each of the SCV classes and reject all other segments. Therefore the classifiers used for spotting should be trained to classify segments belonging to any SCV class and reject all segments that do not belong to SCV classes. For spotting any SCV segment, it is necessary to use a classifier trained for all the SCV classes. It has been pointed out in Chapter 1 that many SCV classes occur infrequently. It is difficult to collect the adequate number of training samples from continuous speech for these classes. This was the reason for limiting the studies in Chapter 3 to a small set of frequently occurring SCV classes. The OCON and ACON classifiers trained with continuous speech data can be used for spotting segments of these classes. A training pattern is derived from speech signal of 100 ms duration with 20 ms before and 80 ms after the VOP in an SCV segment. The patterns that are input to the classifiers during spotting are derived from the segments of 100 ms duration in continuous speech.

Modular neural network architecture has been considered in Chapters 4 and 5 for developing classifiers for all the SCV classes. The classifiers have been trained with isolated utterance data. In order to use these models for spotting , it is necessary to consider the differences in the durations of isolated and continuous speech utterances of SCVs. The isolated utterances of an SCV class are of longer duration compared to the segments of that class in continuous speech. It is observed that most of the necessary information for classification is present in the speech signal of about 100 ms duration around VOP in continuous speech segments. In isolated utterances, the information is present in the speech signal of a longer duration (150 to 200 ms) around VOP. Patterns derived from segments of these durations are used to train the classifiers. For spotting, these classifiers have to be used to classify the patterns derived from 100 ms segments in continuous speech. Two important aspects need to be considered are as follows: (1) Both the patterns should be of same size so that they can be input to a neural network classifier, since isolated speech utterance patterns are used for training and continuous speech patterns are used for spotting, and (2) There should be at least an approximate temporal alignment of the clues in both the patterns. We discuss a method that attempts to take these aspects into account.

We have considered the segment durations of 200 ms, 150 ms and 100 ms for deriving the patterns from isolated utterances. The 200 ms duration consists of 60 ms before VOP and 140 ms after VOP. The 150 ms duration consists of 60 ms before VOP and 90 ms after VOP. The 100 ms duration consists of 20 ms before and 80 ms after VOP. The methods used for deriving patterns containing 20 frames of weighted cepstral coefficients from these different durations have been discussed in Chapter 4. The studies presented in the previous two chapters have shown that the best classification performance is given for the duration of 200 ms. But the duration of 200 ms is not suitable for spotting because many SCV segments in continuous speech are much shorter than 200 ms. The studies on classification have also shown that the performance for the duration of 100 ms is poor. The classifiers trained with the patterns derived from 100 ms segments will give a large number of false alarms during spotting. The classification performance for the duration of 150 ms is close to that for the duration of 200 ms. The patterns for the duration of 150 ms are derived in such a way that the classifiers trained with these patterns can be used for spotting. The speech signal with 60 ms before and 90 ms after VOP is processed to extract 30 frames at a frame rate of 5 ms. The initial 16 frames contain the information about the manner and place of articulation of stop consonants. The remaining 14 frames contain mainly the information about vowel in SCVs. These 14 frames are averaged to obtain 4 smoothed frames from the vowel region. The initial 16 frames along with the 4 smoothed frames are used to form a 20 frame pattern used during training. The patterns that are input during spotting are derived from segments of 150 ms duration in continuous speech using the same method. In section **6.3** we present the results of spotting the frequently occurring SCV classes using the classifiers trained with continuous speech data and the results of spotting all the SCV classes using the classifiers trained with the isolated utterance data.

The classifiers that have been considered so far are trained to classify continuous speech segments or the isolated utterances of SCVs. When these classifiers are used for spotting, in addition to being capable of classifying SCV segments, they should be capable of rejecting segments that do **not** belong to any of the SCV classes. The capability to reject non-SCV segments is important for minimizing the number of false alarms during spotting. Suitable methods for training the classifiers to develop the capability of rejecting non-SCV segments have to be explored.

In the proposed approach for spotting **SCVs,** we first identify the VOPs in continuous speech and then use the classifiers to scan speech segments around the VOPs. In the next section, we discuss a neural network based method for the detection of VOPs in continuous speech.

6.3 Detection of Vowel Onset Points in Continuous Speech

In this section, we develop a neural network **based** method for the detection of vowel onset points in continuous speech. The main aim is to identify the VOPs in SCV segments of continuous speech so that portions of the speech signal around the VOPs can be scanned by the classifiers to hypothesize the presence of SCVs. In Chapter 3, we have developed a method based on the derivative of signal energy for the detection of VOPs in SCV segments manually excised from continuous speech. The VOP in an SCV segment is characterized by a low energy region immediately before and a high energy region immediately after the VOP. Therefore the VOP in an excised segment can be associated with a point at which there is a maximum increase in the signal energy. The VOP can be detected by determining the point at which the energy derivative is maximum. This method cannot be extended for the detection of VOPs in continuous speech. The continuous speech signal of a sentence contains many non-SCV segments which are also characterised with the points at which there is a significant increase in the signal energy. **Therefore** it is necessary to use a more robust method for detecting the VOPs of SCV segments in continuous speech.

Parametric analysis of the speech signal in SCV segments has shown that three parameters, namely, (1) signal energy, **(2)** linear prediction (LP) residual energy and (3) spectral flatness, are significantly different in the regions immediately before and after the VOPs. The signal energy and the LP residual energy parameters show a rapid increase at the VOPs whereas the spectral flatness parameter shows a decrease. These trends in the parameters are illustrated in **Fig.6.1**. The figure shows the plots for speech signal waveform, energy, LP residual energy and spectral flatness parameters. The figure also shows a plot of the derivative of signal energy. This plot shows only the positive portion of the derivative.

We train a multilayer perceptron network to'detect the VOPs by using the trends in the speech signal parameters. The network used for detecting the VOPs consists of **2** hidden layers with 10 nodes in the first hidden layer and 5 nodes in the second hidden layer. The input layer of the network contains 9 nodes and the output layer has 3 nodes. Though the main function of the network is to detect the VOPs, it should also be trained to minimize the number of false alarms in the detection of VOPs. Therefore **3** nodes are used in the output layer of the network with one node (labeled as VOP node) to indicate the presence of VOPs and the other two nodes (labeled as Pre-VOP and Post-VOP) to indicate the absence of VOPs.

For training the network to detect the VOPs, it is necessary to give the values of the above three parameters extracted from a speech signal frame immediately before the VOP and a frame immediately after the VOP. We also give the ratios of the parameters in these two frames because the trend in these parameters contains the information for detection of the VOPs. Thus the signal energy, residual energy and spectral flatness parameters extracted from two frames around the VOP and the ratios of the parameters in the two frames are used to form a vector of 9 values that is input to the network during training. One of the frames starts at 15 ms before

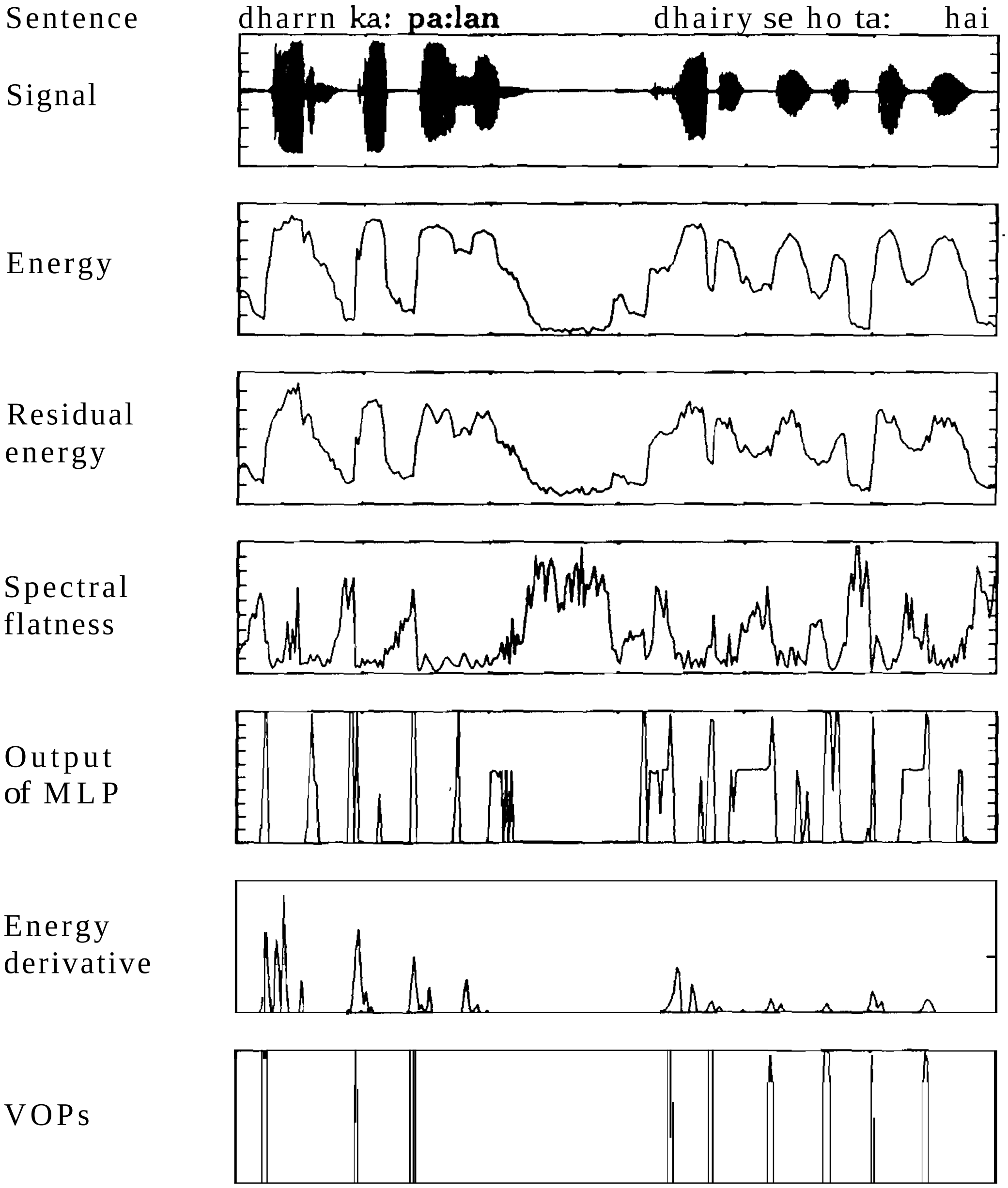


Figure 6.1: Detection of the vowel onset points in the speech signal for the sentence /dharm ka: pa:lan dhairy se hota: hai/. The figure shows the plots of signal waveform, parameters used for detection of the VOPs, output of the multilayer perceptron network, energy derivative and the VOPs detected. It can be seen that the VOPs of all the SCVs are detected.

VOP and the other frame starts at 5ms after VOP. The duration of frames is 10 ms. For training the network, the vectors are extracted from manually excised SCV segments. The desired outputs are specified to give the maximum value (1.O) for the output node,labeled as VOP and the minimum value (0.0) for the other two output nodes labeled as Pre-VOP and Post-VOP. We also extract two other vectors from each SCV segment. One vector is derived from two frames in the region before VOP. One of these frames starts at 35 ms and the other at 15 ms before VOP. The desired outputs for this training vector specify the maximum value for the Pre-VOP node and minimum values for the VOP and Post-VOP nodes. Another vector is derived from two frames in the region after VOP. One of these frames starts at 35 ms and the other at 55 ms after VOP. The desired outputs for this training vector specify the maximum value for the Post-VOP node and minimum values for the VOP and Pre-VOP nodes. The network is trained for about 150 SCV segments excised from continuous speech.

For the detection of VOPs in continuous speech using the network trained as above, a 9-dimensional parameter vector extracted at every 10 ms is given as input to the network and the output of the network indicates the presence or absence of VOP at that point in continuous speech. The parameter vector is extracted from two frames with one frame starting at the point under consideration and another frame starting at 20ms after this point. Thus the continuous speech signal of a sentence is scanned by the network to detect the VOPs. Fig.6.1 shows the output of the network for the sentence whose speech signal waveform and parameters are also plotted in the figure. The network detects the VOPs of all the SCV segments in the sentence. But it also hypothesizes many other points as VOPs. It is interesting to note from the plots of the network output and the energy derivative that both of them have peaks at the VOPs of SCVs. At all other points only one of them shows a high value or a peak. We combine the network output and the energy derivative to eliminate most of the spurious hypotheses. The VOPs detected in the speech signal for the sentence /dharm ka: pa:lan dhairy se hota: hai/ using this method are shown in Fig.6.1. They include the VOPs in SCVs (/dha/, /ka:/, /pa:/, /dhai/, /ta:/) and also the VOPs of the segments belonging to non-SCV classes (/ry/, /se/, /ho/, /hai/). The objective of our method is to detect the VOPs in SCV segments so that scanning of speech signal during spotting can be limited to only these segments. This method has been used for detection of the VOPs for 50 sentences and the results have shown that the VOPs of the SCV segments are detected with nearly 100% accuracy. The method has failed to detect the VOPs in a few segments where the signal energy is low. It is observed that the method also detects the VOPs in many CV segments belonging to non-SCV classes. The VOPs of most of the fricative CV segments have been detected. But the VOPs of many CV segments where the signal energy in the consonant regions is high are not detected. These segments mainly belong to the nasal (/m/ and /n/), semivowel (/y/ and /w/) and lateral (/l/) consonants. For example, the VOP of the /la/ segment in the sentence of Fig.6.1 is not detected. In the next section, we study the effects of the detection of VOPs on spotting performance.

6.4 Studies on Spotting SCVs

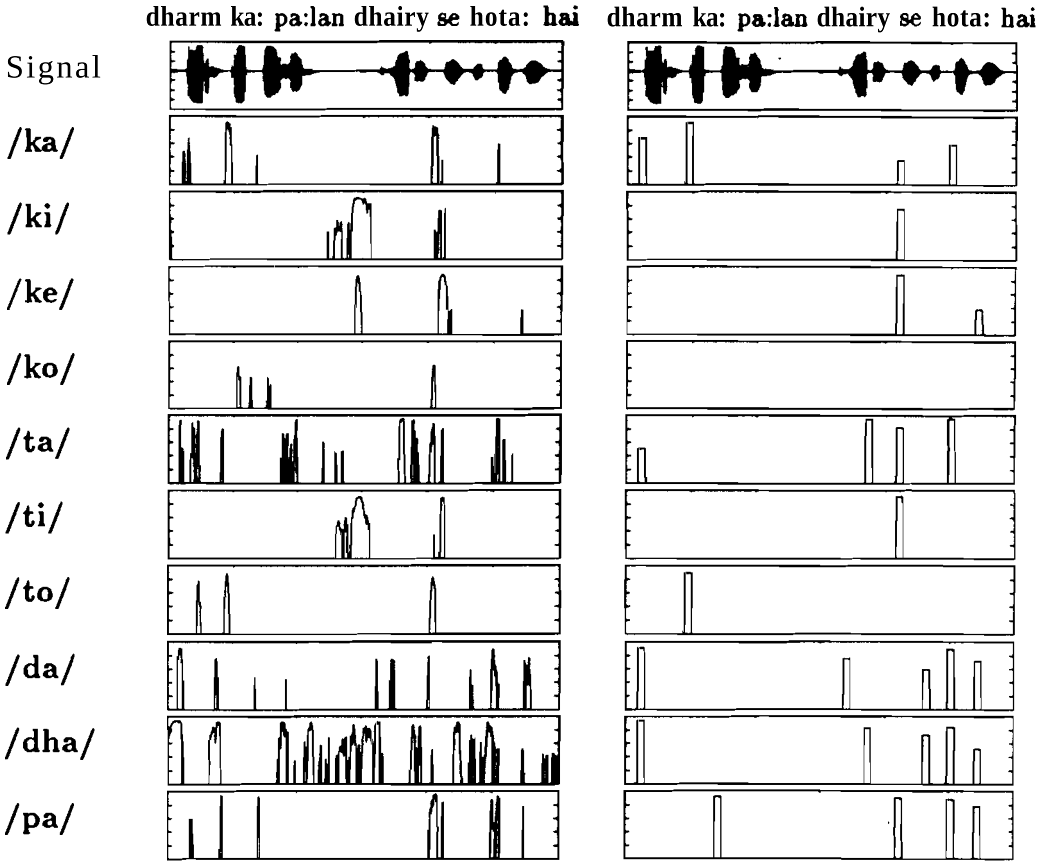
In this section, we present the results of our proposed approach for spotting. We first present the results of spotting frequently occurring SCV classes using the classifier models trained with the segments excised from continuous speech. Then we present the results of spotting all the SCV classes using the constraint satisfaction model developed for classification of isolated utterances.

6.4.1 Spotting using Classifiers Trained with Continuous Speech Data

In Chapter 3, we have developed the OCON and ACON classifiers for the set of ten frequently occurring SCV classes. In this section, we study the performance of these classifiers in spotting the SCV segments in continuous speech.

We first study the effect of the detection of VOPs on spotting performance. For this study, we consider the OCON classifier for the following ten classes: /ka/, /ki/, /ke/, /ko/, /ta/, /ti/, /to/, /da/, /dha/ and /pa/. This classifier contains one multilayer perceptron network trained for each class. The classifier is first used to scan the speech signal of a sentence continuously. For scanning, a pattern is derived at every 5 ms from the signal of 100 ms duration and the pattern is input to the networks. The outputs of the networks for the ten classes are plotted in Fig.6.2(a) for a Hindi sentence. It can be seen from the figure that though the presence of the segments belonging to each of the ten classes is indicated, the number of false alarms given by the networks is large.

We have used the same classifier for spotting using our approach in which the VOPs are detected first. Once the VOPs are detected, a segment around each VOP is scanned by the classifier. The scanning window around a VOP includes 50 msec before the VOP and 100 msec after the VOP. We consider a scanning window around VOP for spotting to take into account any ambiguity in the detection of VOPs precisely. It is thus enough if the method gives an approximate location of the VOP of an SCV segment in continuous speech. The pattern derived from a 100 msec long speech signal at every 5 msec in the scanning window is input to the classifier networks. The maximum output value in a scanning window for a class is assigned as the output of the network for that class. The outputs of the networks for the ten classes using this approach are plotted in Fig.6.2(b). It can be seen from Figs.6.2(a) and (b) that many false alarms given in the standard approach for spotting are eliminated using



(a) Without detection of VOPs

(b) With detection of VOPs

Figure 6.2: Spotting SCV segments in continuous speech of the sentence /dharm ka: pa:lan dhairy se hota: hai/ using the classifier based on the OCON architecture and MLP model trained for ten frequently occurring SCV classes. The plots show the outputs of the networks for different classes in spotting SCVs (a) Without detection of VOPs and (b) With detection of VOPs. It can be noted that the number of false alarms is much smaller for spotting with detection of VOPs.

our approach. This has been possible mainly due to limiting the scanning by the classifier: to segments around the VOPs. Another advantage of our approach is that the computational complexity is reduced significantly by limiting the scanning.

We have studied the performance of spotting using different classifiers. The outputs of different classifiers used for spotting the SCVs in the continuous speech signal of a Hindi sentence are shown in Figs.6.3(a) and (b). These figures show the speech signal waveform for the sentence and the VOPs detected. These figures also show the outputs of the classifiers for each class in the scanning windows around the VOPs. The outputs of the OCON classifier are shown in Fig.6.3(a) and the outputs of the ACON classifier in Fig.6.3(b). It is observed that the presence of the segments belonging to the ten SCV classes is indicated by both the classifiers. The ACON classifier gives fewer false alarms than the OCON classifier. The false alarms in segments belonging to non-SCV classes are mainly due to inadequate training of the classifiers in rejecting them.

It is observed that for each segment around a VOP, presence of more than one SCV class is hypothesized. It is interesting to note that the hypotheses include mainly the classes that are phonetically close to the actual class of the segment. We consider the hypotheses made by the ACON classifier in Fig.6.3(b) for illustration. The presence of the classes /ka/ and /pa/ is hypothesized for the /ka:/ segment. The output values of the classifier can be considered as evidences for the classes. It is noted that the evidence for the class /ka/ is stronger than the evidence for the class /pa/: A similar performance has been observed for the segments of /pa:/ and /ta:/. For the /dha/ segment, the hypotheses include the classes /da/ and /pa/ in addition to the class /dha/. The evidence for the class /pa/ is slightly stronger than the evidence for /dha/. A postprocessor can be used to process the outputs of the classifier to determine the class of a segment using the evidences available. The postprocessor can use a simple

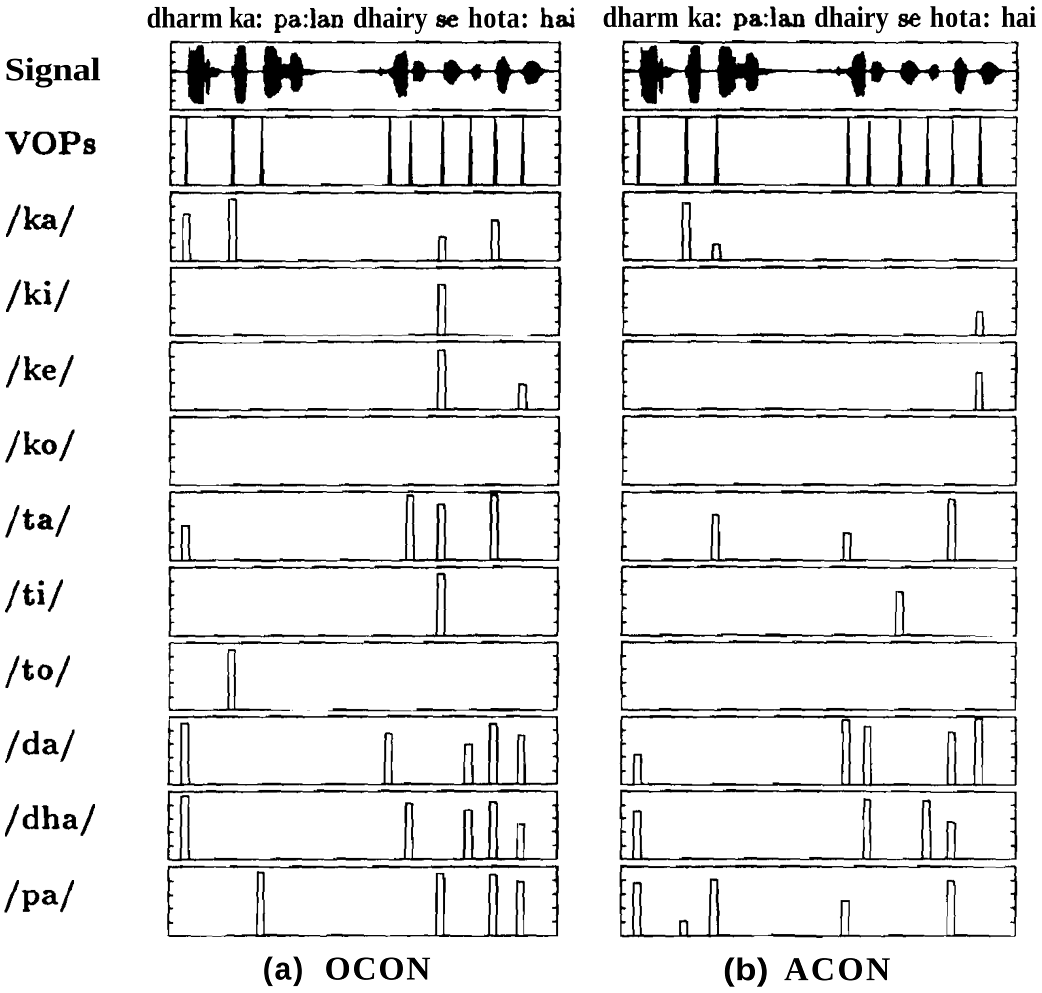


Figure 6.3: Spotting SCV segments in continuous speech of the sentence /dharm ka: pa:lan dhairy se hota: hai/ using the MLP based classifiers trained for ten frequently occurring SCV classes. The plots show the outputs of the networks in the classifiers based on (a) OCON architecture and (b) ACON architecture.

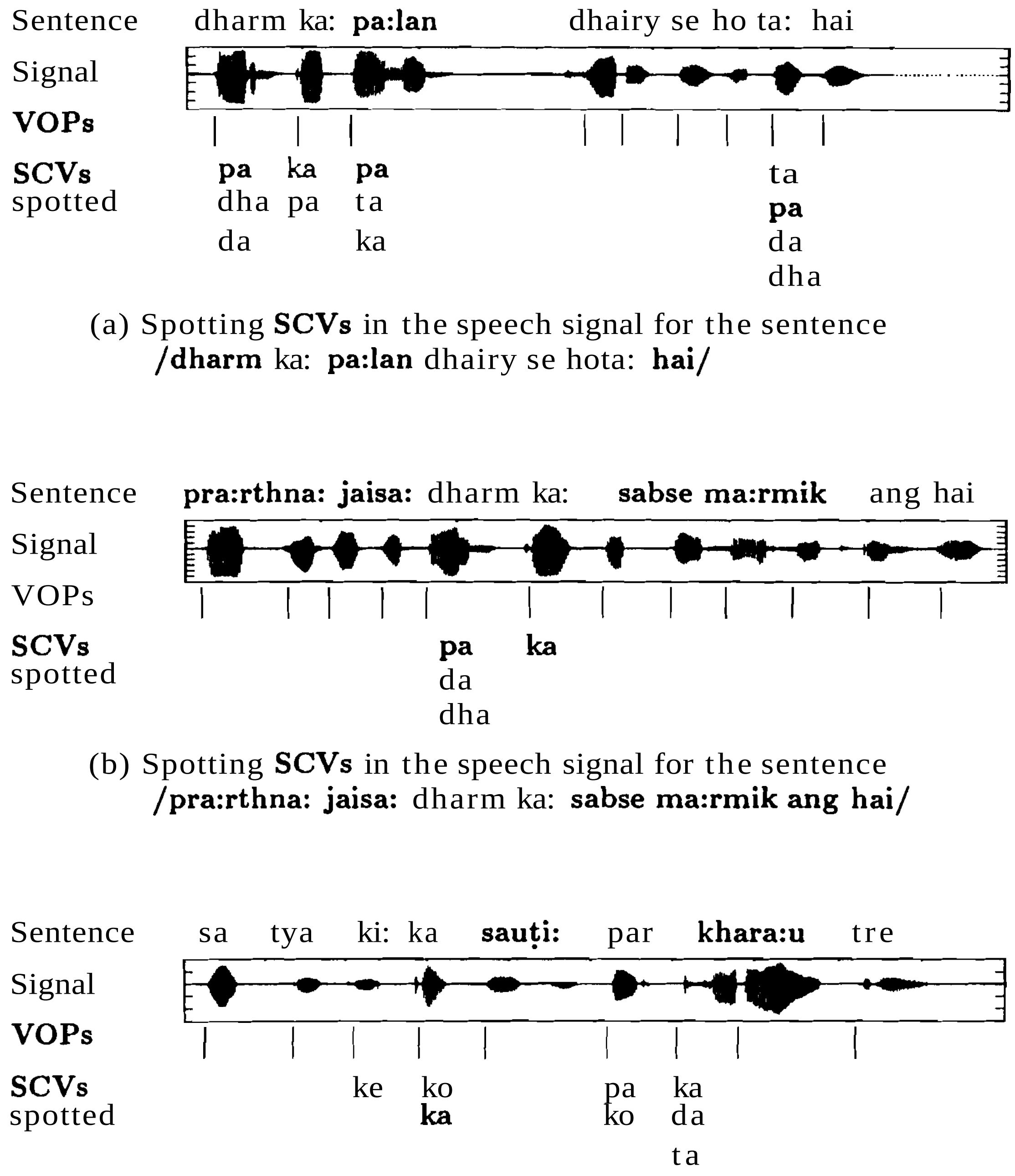
method of assigning the class with the strongest evidence to a segment. If this method is used, the class /pa/ will be assigned to the /dha/ segment. A better method is to use the knowledge of similarities among the classes as constraints in combining the evidences for the classes. It has been shown in the previous chapter that this method used in a constraint satisfaction model has significantly improved the performance for classification of SCV utterances.

The classes hypothesized by the ACON classifier for different SCV segments and rank ordered based on the evidences available are given for six sentences in Figs.6.4 and Fig.6.5. The text of the sentence, speech signal waveform, vowel onset points detected and the classes hypothesized for the SCV segments are given for each sentence. It can be seen from these figures that the VOPs of most of the SCV segments have been detected. The VOPs of a few SCV segments occurring mainly at the end of the phrases and sentences, and characterized by low signal energy have not been detected. For example, the VOPs of the /ṭi:/ segment in Fig.6.4(c) and the /bhi:/ segment in Fig.6.5(a) have not been detected.

It is observed that for some of the SCV segments, the actual class is not present among the classes hypothesized. For example, the hypotheses for the /ki:/ segment in Fig.6.4(c), the /ti/ segment in Fig.6.5(a), and some of the /ke/ segments in Figs.6.5(b) and (c) do not include the actual class. But the hypotheses include the classes that are phonetically close to the actual class.

It is also observed that the classes that are phonetically close to the actual class are included in the hypotheses for the segments of the SCV classes for which the classifier has not been trained. For example, the hypotheses for the /kha/ segment in Fig.6.4(c) include the class /ka/, and the class /dha/ has been hypothesized for the /bha:/ segment in Fig.6.5(a).

The above results show that the classifiers trained with the patterns derived from

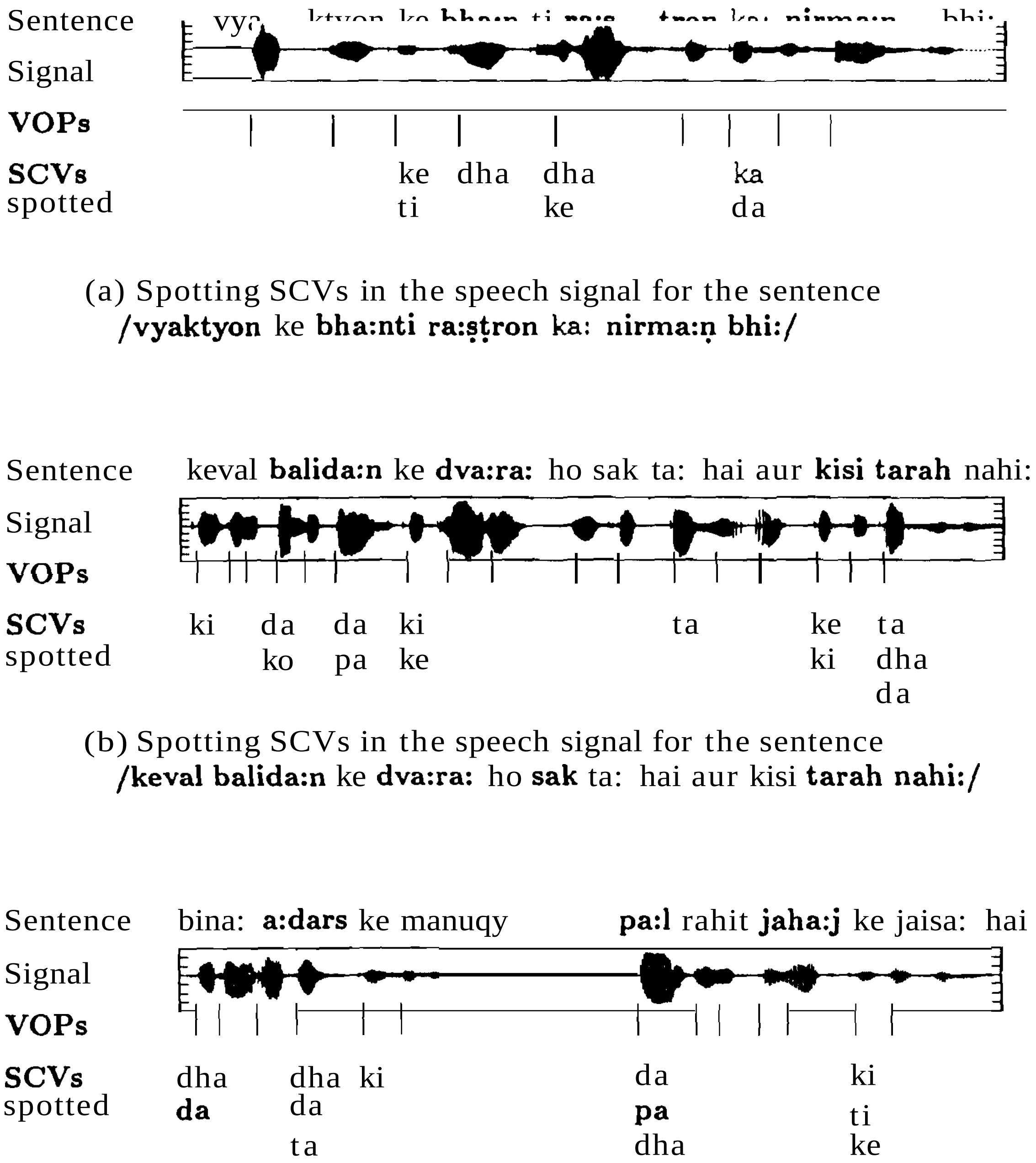


(a) Spotting SCVs in the speech signal for the sentence /dharm ka: pa:lan dhairy se hota: hai/

(b) Spotting SCVs in the speech signal for the sentence /pra:rthna: jaisa: dharm ka: sabse ma:rmik ang hai/

(c) Spotting SCVs in the speech signal for the sentence /satya ki: kasauṭi: par khara: utre/

Figure 6.4: Spotting SCVs in the speech signal for different sentences using the classifier based on ACON architecture and MLP model trained with continuous speech data for ten SCV classes. The figure for a sentence gives the text, signal waveform, vowel onset points (VOPs) identified and SCV segments spotted. The hypotheses for non-SCV segments are not indicated.



(a) Spotting SCVs in the speech signal for the sentence /vyaktyon ke bha:nti ra:ṣṭron ka: nirma:ṇ bhi:/

(b) Spotting SCVs in the speech signal for the sentence /keval balida:n ke dva:ra: ho sak ta: hai aur kisi tarah nahi:/

(c) Spotting SCVs in the speech signal for the sentence /bina: a:dars ke manuqy pa:l rahit jaha:j ke jaisa: hai/

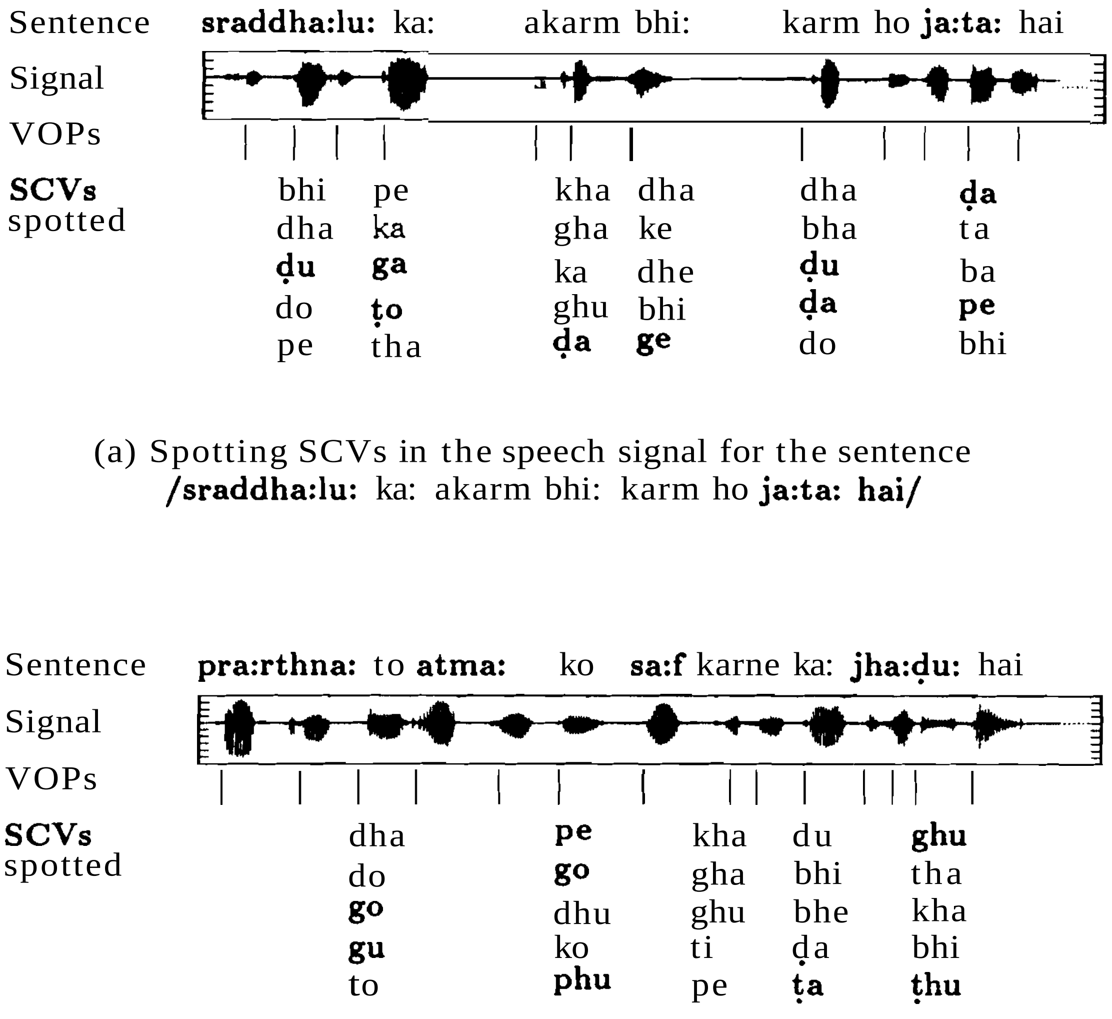
Figure 6.5: Spotting SCVs in the speech signal for different sentences using the classifier based on ACON architecture and MLP model trained with continuous speech data for ten SCV classes. The figure for a sentence gives the text, signal waveform, vowel onset points (VOPs) identified and SCV segments spotted. The hypotheses for non-SCV segments are not indicated.

manually excised segments can be used for spotting. The performance of our proposed approach for spotting depends on the performance of the method used for the detection of VOPs and the performance of the classifiers. This approach can be extended for spotting the segments of all the SCV classes.

6.4.2 Spotting using Constraint Satisfaction Model

In the previous chapter, we have developed a constraint satisfaction model (CSM) for classification of all the SCV classes. The model has been trained with the patterns derived from the isolated utterances of SCVs. We have used this model as the classifier for spotting the segments of all the SCV classes in continuous speech. After the detection of the VOPs in the speech signal of a sentence, a segment of 150 ms duration around each VOP is processed to derive a pattern as explained in Section 6.2. This pattern is input to the CSM to determine its class. The classes of the five largest values among the output values of the instance pool units in the CSM are considered as the hypotheses for each segment. The output values for the classes are considered as evidences for the hypotheses. The spotting performance of the CSM is illustrated for two sentences in Fig.6.6. The text of the sentence, speech signal waveform, vowel onset points detected and the five hypotheses for each the SCV segments are given for each sentence. It is seen that the hypotheses for many SCV segments include the actual class. For some segments, the hypotheses include the classes that are phonetically close to the actual class.

The above result shows the potential of our approach for spotting in using it for all the 80 SCV classes. The main difficulty is in matching the training patterns derived from isolated utterances with the patterns derived from continuous speech segments during spotting. A significant improvement in the spotting performance can be expected for a CSM trained with continuous speech SCV data.



(a) Spotting SCVs in the speech signal for the sentence /sraddha:lu: ka: akarm bhi: karm ho ja:ta: hai/

(b) Spotting SCVs in the speech signal for the sentence /pra:rthna: to atma: ko sa:f karne ka: jha:ḍu: hai/

Figure 6.6: Spotting SCVs in the speech signal for different sentences using the constraint satisfaction model developed for classification of isolated utterances of all the SCV classes. The figure for a sentence gives the text, signal waveform, vowel onset points (VOPs) identified and the first 5 class hypotheses for SCV segments spotted. The hypotheses for non-SCV segments are not indicated.

6.5 Summary and Conclusions

A summary of the issues addressed and studies carried out for spotting the SCV segments in continuous speech is given in Table-6.1.

Table 6.1: Summary of the studies on spotting the SCV segments in continuous speech.

1. An approach based on the detection of vowel onset points and scanning around them for spotting SCVs in continuous speech has been developed.
2. A neural network based method for the detection of vowel onset points in continuous speech has been developed.
3. The spotting performance of classifiers trained with the continuous speech data for a subset of the SCV classes has been illustrated.
4. The spotting performance of the constraint satisfaction model trained with isolated utterance data for all the SCV classes has been illustrated.
5. The spotting approach based on the detection of vowel onset points has been shown to give fewer false alarms than the standard approach for spotting where the speech signal of a sentence is completely scanned.

In this chapter, we have discussed the main issues in spotting the SCV segments. We have proposed an approach in which the vowel onset points are detected first and the segments around these points are scanned by classifiers. We have illustrated the results of spotting ten frequently occurring SCV classes using the classifiers trained with continuous speech data. We have shown that our approach eliminates many false alarms given by the standard approach for spotting. We have also illustrated the performance of the constraint satisfaction model developed for classification of isolated utterances in spotting segments of all the SCV classes. It is necessary to evolve suitable techniques for using the classifiers trained with isolated utterance data for spotting subword units in continuous speech. It is also important to explore learning algorithms that minimize spotting errors [48] for improving the performance of systems for spotting subword units.

The focus of our studies so far have been on developing models for classification and spotting of SCV segments. The results of the studies on classification have shown that a significant percentage of errors is due to misclassification of the place of articulation of stop consonants. The patterns derived for training and testing the classifiers have been based on the weighted cepstral coefficient representation for the speech signal of the SCV segments. In the next chapter, we explore suitable representations for the transition regions in the SCV segments in order to improve the performance for classification of the place of articulation.

Chapter 7

PARAMETRIC REPRESENTATIONS FOR SCV TRANSITIONS

7.1 Need for Classification of SCV Transitions

Our studies on classification of SCVs in Chapters **3** and 4 have shown that a significant percentage of errors is due to misclassification of the place of articulation of the stop consonants. Parametric representation based on weighted cepstral coefficients has been used in these studies. In this chapter, we explore methods for suitable representation of transitions in SCVs which contain the information for classification of the place of articulation.

The important clues for identifying the place of articulation of stop consonants are: (1) characteristics of the spectrum during the release of the stop consonant (burst event) and **(2)** formant transitions between the stop consonant and its following vowel (transition event). These clues are dependent on the adjacent vowels [10] [83]. Therefore, it is difficult to identify the POA without using the contextual vowel information. It may be easier to identify the POA along with the context. That is, instead of having 'Velar', 'Alveolar', 'Dental' and 'Bilabial' as the classes, it may be advantageous to have classes such as 'Velar in the context of vowel /a/ (Velar_a)'. We attempt to develop a classifier for POA in the context of a vowel. By having a separate class for POA in the context of each vowel, what we are trying to do is to classify the type of SCV transitions [70]. Our focus is on the design of a classifier for the SCV transitions that occur in continuous speech.

7.2 Issues in Classification of SCV Transitions

For the design of a classifier for SCV transitions, one has to determine the portion of the speech signal to be analysed. As the important clues for identification of the SCV transition are the release spectrum and the formant transitions, it is necessary that the analysis region includes the speech signal between the beginning of the release of the stop consonant and the steady formant region of the following vowel. The beginning of the release can be identified, at least manually. But it is difficult to identify the point at which the vowel formants have become steady. This is more so in continuous speech, because vowels tend to have short steady formant regions. As an approximation, we consider the portion of the speech signal of a fixed duration (about 40 ms) from the beginning of the release instant as the transition region.

Processing of speech signal for classifying the type of SCV transition should aim at capturing the characteristics of the release spectrum and the transition. Here, we focus on the analysis of the transition. The spectral characteristics vary during the transition region. Suitable representation of the dynamic characteristics of the transition region is important for recognition of the stop consonants [12]. Time varying features were identified as important cues for perception of the place of articulation of the stop consonants [84]. Differences in formant frequencies at the beginning and end of the transition region and the duration of the transition region were used for recognition of the place of articulation of unaspirated stop consonants in Telugu [85]. Processing methods using analysis based on time varying, data selective models of the speech signal in VC transitions have been shown to give improved performance compared to the standard processing methods [69]. In our study, we use parametric representations based on standard processing methods and also propose parametric representations extracted using pitch region analysis to capture the characteristics of the transitions in SCV segments.

7.3 Classes of SCV Transitions

Stop consonants are characterized by their manner of articulation and place of articulation. There are four different places of articulation and four different manners of articulation. The formant transitions are similar for the stop consonants with the same POA in the context of a particular vowel. Therefore the SCV transitions of consonants with the same POA but with different manners of articulation are grouped into a single class. Moreover, the formant transitions for a particular POA in the context of short and long versions of a vowel (such as /a/ and /a:/) are not much different. Taking these observations into consideration, the following classes of SCV transitions were selected for our study.

(1) Velara
(2) Velari
(3) Velar-u
(4) Velar_e
(5) Velar_o
(6) Alveolara
(7) Alveolari
(8) Alveolar-u
(9) Alveolar2
(10) Alveolar_o
(11) Dental_a
(12) Dental_i
(13) Dental-u
(14) Dental-e
(15) Dental-o
(16) Bilabial_a
(17) Bilabiali
(18) Bilabial-u
(19) Bilabial-e
(20) Bilabial-o

Only 13 of these classes were considered here as there was not enough speech data available for the other 7 classes (5 classes with Alveolar as POA, Bilabial2 and Bilabial_o).

7.4 Parametric Representations

In our approach to the classification of the SCV transition type, we have used eight methods to represent the characteristics of the transition region. Four processing methods use an arbitrarily placed window of a fixed frame size. These methods are not well suited for capturing the spectral changes in the transition region, which are important for classifying the SCV transition type. This is because the changes in the vocal tract system during the analysis interval and the window effects mask the slow changes in the vocal tract shape during the transition region. For classification, it is necessary to capture the spectral changes in the transition region. The spectral information in the high energy portion of each pitch cycle is likely to provide reliable spectral information. The high energy portion is usually around the significant point of excitation in the pitch cycle. Therefore, an analysis window centered around the high energy region of the pitch cycle is used. We propose pitch region analysis to obtain a robust parametric representation for the SCV transitions. We briefly discuss implementation details of the processing methods used in our study.

Formant Frequencies and Amplitudes (FMT)

The first three formants were extracted using group delay processing of signal [86] for four frames starting at the vowel beginning point in an SCV transition region. The three formant frequencies and their amplitudes for four frames were normalized and used to form a 24-dimensional vector for each SCV transition region.

Spectral Coefficients (SPC)

Spectral coefficients were extracted for four frames starting at the vowel beginning point in an SCV transition region. 16 mel-scale spectral coefficients were computed using a 128-point FFT with a frame size of 12.8 ms and an overlap of 9.6 ms between successive frames. The normalized spectral coefficients were used to form a 64-dimensional vector for each SCV transition region.

Mel-scale Weighted Cepstral Coefficients (MCEP)

32 mel-scale spectral coefficients were computed using a 512-point FFT with a frame size of 12.8 ms and an overlap of 9.6 ms between successive frames. 10 weighted cepstral coefficients were derived from these mel-scale spectral coefficients for each of

the first four frames in the transition region. These cepstral coefficients were used to form a 40-dimensional vector for each SCV transition region.

LP Derived Weighted Cepstral Coefficients (WCEP)

Ten weighted cepstral coefficients derived from linear prediction coefficients were extracted from each of the first four frames in an SCV transition region. Thus a 40-dimensional vector was obtained for each SCV transition.

Pitch Region Analysis based Parameters

In order to extract the parametric representation from the high energy region of each pitch cycle, it is necessary to determine, at least approximately, the significant points of excitation in the pitch cycles. We have used a method based on the positions of the peaks in the energy curve of a low-pass filtered (with 1 KHz cut-off) linear prediction residual signal. These peaks indicate the significant points of excitations in the pitch cycles. For each SCV transition region, significant points of excitation in the first four pitch cycles in the beginning of the transition region were determined. A pitch region is defined as the region with duration of the pitch period and centered around the significant point of excitation in the pitch cycle. For every such pitch region, the formant frequencies and amplitudes (PRFMT), spectral coefficients (PRSPC), mel-scale weighted cepstral coefficients (PRMCEP) and LP derived weighted cepstral coefficients (PRWCEP) were extracted.

7.5 Studies on Classification of SCV Transitions

7.5.1 Data Collection and Preparation

Speech data for SCV transitions was collected from 50 sentences for three male speakers and three female speakers. The regions of analysis were manually identified in the digitized speech signal. The duration of the analysis region was fixed at 40 ms, with 10 ms before and 30 ms during the transition region. This duration was arrived at from the observation of the durations of SCV transition regions on a spectrograph. Data for 300 SCV transition regions was collected for each of the six speakers.

7.5.2 Classification Model

A multilayer perceptron model with two hidden layers was used as a classifier in our studies. The number of input nodes depends on the parametric representation used. The number of nodes in the first and second hidden layers were chosen as 30 and 20, respectively. The number of output nodes were 13, representing the 13 classes used.

7.5.3 Classification Studies and Results

Speaker Dependent (SD) Classification

In this study, different multilayer perceptron networks were trained for each of six speakers. 30% of the total data available for a speaker was used in training the network. The remaining 70% of the total data was used as the test data. This study was performed for each of the eight different parametric representations. The performance for different parametric representations is given in Table-7.1.

Multispeaker (MS) and Speaker Independent (SI) Classification

In this study, a network was trained with 30% of the data from four (two male and two female) speakers. The remaining 70% of the total data (853 patterns) for these four speakers was used as multispeaker test data. The total data (568 patterns) of the other two speakers was used as speaker independent test data. The performance for different parametric representations is given in Table-7.1.

Table 7.1: Comparison of the performance of different parametric representations for classification of SCV transitions. The performance is given for two cases of classification criterion. Case-1 refers to the criterion that the class of a given SCV transition pattern is the class of the largest value among the outputs of the classifier. Case-2 refers to the criterion that the class of a given SCV transition pattern is amongst the classes of the largest and the second largest output values of the classifier. The performance for each parametric representation is given on data sets of speaker dependent (SD), multispeaker (MS) and speaker independent (SI) classification. The entries in the parantheses indicate the total number of patterns in a data set. It can be seen that the pitch region analysis based formants (PRFMT) and spectral coefficients (PRSPC) give a better performance compared to the fixed frame size analysis based formants (FMT) and spectral coefficients (SPC) respectively.

(a) Performance for Case-1 of classification criterion.

Parametric	Training Data		Test Data		
Representation	SD	MS	SD	MS	SI
	(523)	(348)	(1246)	(853)	(568)
FMT	84.1	60.6	33.7	29.8	27.0
SPC	92.0	85.9	43.5	43.5	34.0
MCEP	94.5	78.7	45.6	50.5	34.9
WCEP	99.6	97.4	65.1	58.6	40.3
PRFMT	92.7	76.7	39.2	34.6	34.7
PRSPC	97.7	95.7	52.2	48.3	37.8
PRMCEP	93.7	72.1	42.0	31.6	27.5
PRWCEP	98.7	93.4	51.6	49.3	39.9

(b) Performance for Case_2 of classification criterion.

Parametric	Training Data		Test Data		
Representation	SD	MS	SD	MS	SI
	(523)	(348)	(1246)	(853)	(568)
FMT	89.5	73.6	56.8	51.3	46.8
SPC	95.4	92.2	65.7	65.3	54.2
MCEP	95.6	87.9	65.2	70.6	57.2
WCEP	99.8	97.7	81.4	78.0	58.6
PRFMT	92.9	85.1	58.3	60.6	54.7
PRSPC	97.9	95.7	71.8	68.7	56.7
PRMCEP	95.3	83.9	62.3	57.2	49.5
PRWCEP	99.0	93.7	72.5	68.5	59.9

7.5.4 Analysis of Performance for SCV Transitions

The results indicate that for formants and spectral coefficient representations, the pitch region analysis based processing methods give a better classification performance compared to the standard processing methods. For the cepstral coefficient representations, the standard processing methods give a better performance. This is because of the short duration segments used in the pitch region analysis.

The best performance on training data for speaker dependent classification is above 99% and that for multispeaker classification is above 97%. The performance reported in [85] for classification of the place of articulation of unaspirated stop consonants in the context of known vowel and using the data collected from isolated words for **3** speakers is about 70%.

On multispeaker test data, the best performance obtained in our studies is about 58% for Case-1 and is about 78% for Case_2. On speaker independent test data, the best performance obtained is about 40% for Case-1 and is about 60% for Case_2. Considering the large number of SCV transition classes used and confusability amongst the classes, this accuracy is significantly high. Human performance is also not likely to be high for this data.

It is observed that the improvement in the performance for the pitch region analysis based methods for spectral coefficient representation is significantly higher for the male speaker data compared with that for the female speaker data, as given in Table-7.2. The poor performance for the female speaker data can be due to the short (about 3.0 ms) analysis window used in the pitch region analysis based methods. It is necessary to explore better methods of processing using short analysis windows to improve the performance for the female speaker data.

Table 7.2: Comparison of performance in classification of SCV transitions for male speaker data and female speaker data. The pitch region analysis based method gives a better performance for male speaker data and a poorer performance for female speaker data compared to the fixed frame size analysis based method.

Test data	Male Speaker Data		Female Speaker Data	
	SPC	PRSPC	SPC	PRSPC
Speaker Dependent	42.4	57.0	45.5	49.2
Multispeaker	40.8	56.2	46.2	40.4
Speaker Independent	36.1	45.6	32.1	29.9

7.6 Fuzzy Nature of Clues for SCV Transitions

In continuous speech the same SCV may occur in different contexts. Therefore there may be variability in the features of the utterance due to variability in speech production as well as due to context. Moreover, there will also be variability in speech production due to different speakers. All these factors lead to feature data that can best be described in linguistic terms, such as 'low', 'medium' and 'high', which in turn can best be expressed as values of membership functions of fuzzy sets.

It is necessary to represent the production information in the speech signal in suitable parameters or features for input to a classifier. Parameters like spectral coefficients and cepstral coefficients are likely to be influenced by the nature of signal processing as well, besides the natural variations in the production process. Variations due to signal processing operations contribute to distortion and noise, rather than fuzziness. Therefore it is preferable to consider articulatory or related acoustic parameters like formants as features representing the SCV transitions. Formants are relatively easier to extract compared to the articulatory parameters. Formant features also reflect the dynamics of the vocal tract system in the form of formant trajectories. Therefore the formants were selected as parameters to represent the SCV transitions in this study.

Speaker variability is caused due to differences in the dimensions of the vocal tract systems. In order to compensate this to some extent, ratios of formants may be considered as features. Since we are considering in this study only data from two speakers, we have decided to consider only the formant values as features. Formant data is collected for successive frames of speech signal data in each SCV transition,

Formants are resonances of the vocal tract system, and hence any natural variations in the shape of the vocal tract are reflected in these resonances as well. Since variability due to speech, context and speaker are all preserved in the formant trajectories, the formant data can be assumed fuzzy, and the data is fuzzified before feeding it to a neural network classifier for training and testing.

Fuzzification of formant data involves several issues. For example, one could fuzzify the features individually in the frequency and time domains. But it appears more logical if the fuzzification could be done knowing that the three formants should occur together as a set in each frame. Also the formants in successive frames are not independent. Hence this dependency should also be considered in fuzzifying the input data to the neural network classifier.

It is natural to expect that the class labels will not be crisp either, due to significant overlap of features across the different classes of SCV transitions. Therefore, for effective classification, it is preferable that the output classes are fuzzy. In the next section we describe a fuzzy neural network classifier that takes fuzzy input data.

7.7 Fuzzy Neural Network Classifier

It was shown in [38] that fuzzification of input data and the output class label data improves the classification performance of a multilayer perceptron network for recognition of vowels using formants as features. The network takes as input the values of fuzzy membership functions for each of the three formants. Each input feature F_j in quantitative form is expressed in terms of membership values to each of the three linguistic properties 'Low', 'Medium' and 'High'. The π membership function is used to assign membership values for the input features. The π membership function in one-dimensional form, with range $[0,1]$, is defined as given below.

$$\pi(x : c, r) = \begin{cases} 2(1 - (\frac{|x-c|}{r}))^2, & \text{f} \quad \frac{r}{2} < |x-c| < r, \\ 1 - 2(\frac{|x-c|}{r})^2, & \text{for} \quad 0 < |x-c| < \frac{r}{2}, \\ 0, & \text{otherwise}, \end{cases} \tag{7.1}$$

where x is a pattern point, r is the radius of the π function and c is the central point.

The fuzzy sets for the linguistic properties 'Low', 'Medium' and 'High' for each formant are represented by membership functions π_L, π_M and π_H respectively. The parameters of these membership functions are defined below.

Let F_{jmax} and F_{jmin} be the upper and lower bounds of feature F_j in all pattern points. For the three linguistic property sets, parameters are defined as

$$\begin{aligned} r_M(F_j) &= 0.5 * (F_{jmax} - F_{jmin}) \\ c_M(F_j) &=. \; F_{jmin} + r_M(F_j) \\ r_L(F_j) &= \frac{(c_M(F_j) - F_{jmin})}{fdenom} \\ c_L(F_j) &= c_M(F_j) - 0.5 * r_L(F_j) \\ r_H(F_j) &= \frac{(F_{jmax} - c_M(F_j))}{fdenom} \\ c_H(F_j) &= c_M(F_j) + 0.5 * r_H(F_j) \end{aligned} \tag{7.2}$$

where '$fdenom$' is a parameter controlling the extent of overlapping.

The three π membership functions are defined for each of the three formants and for each of the N frames in the transition region. Thus a SCV transition is represented by an Nx9-dimensional matrix of membership values. Such Nx9-dimensional patterns are used as input to a neural network classifier.

During the training phase, the desired output vector is expressed as the desired membership values, lying in the range [0,1]. To obtain these membership values, the distance of a training pattern F from the average pattern O_k for the kth class is defined as

$$z_k = \sqrt{\frac{1}{9N}\sum_{i=1}^{N}\sum_{j=1}^{9}(F(i,j) - O_k(i,j))^2} \tag{7.3}$$

The membership value for the training pattern F to the kth class is defined as

$$d_k(F) = \frac{1}{1 + (\frac{z_k}{f_d})^{f_e}} \tag{7.4}$$

where the positive constants f_d and f_e control the amount of fuzziness in the class-membership set. The desired output vector for a training pattern is obtained by computing the membership values for the pattern to each of the classes and used in training the multilayer perceptron network.

In fuzzification of the input data, the formant features for each frame are fuzzified independently. But, there is a sequence of frames in each SCV transition, and the data in each frame depends to some extent on the adjacent frames. This fact must be used in fuzzification of formant trajectories. Two methods of fuzzification of sequences of formant data are presented in the next section.

7.8 Fuzzification of Formant Trajectories

The formant data for one frame is dependent on the adjacent frames. This time-dependency can be incorporated in the fuzzification of the trajectories by reducing the variability allowed for the subsequent frames given the variability of the current frame. The reduction in variability allowed for subsequent frames can be realized by decreasing the radii of the π membership functions for fuzzy subsets of features in those frames, and correspondingly modifying the centers of the functions.

The parameters of the membership functions for the features in the first frame are defined as explained in the previous section. The parameters of the functions for subsequent frames are obtained from those of the first frame as follows:

$$\begin{aligned}
r_{iM}(F_j) = &\ (1-a)^{*} r_{(i-1)M}(F_j) \\
c_{iM}(F_j) = &\ F_{jimin} + B^{*} r_{iM}(F_j) \\
r_{iL}(F_j) = &\ (1-a) * r_{(i-1)L}(F_j) \\
c_{iL}(F_j) = &\ c_{iM} - 0.5^{*} \mathrm{B}^{*} r_{iL}(F_j) \\
r_{iH}(F_j) = &\ (1-a)^{*} r_{(i-1)H}(F_j) \\
c_{iH}(F_j) = &\ c_{iM} + 0.5 * \mathrm{B}^{*} r_{iH}(F_j)
\end{aligned} \tag{7.5}$$

where i is the frame number and $2 < i < \mathrm{N}$. The constants a and B are chosen such that the distance between the average patterns of the classes is maximum. Typical values for the constants a and B are 0.075 and 1.30, respectively.

Another way of incorporating the time dependency is to use multidimensional membership functions for groups of adjacent frames. The parameters for the multidimensional membership functions are obtained from the parameters of the one-dimensional membership functions of features for individual frames. The definition of one-dimensional π function in equation 7.1 is extended for an n-dimensional π function of a group of n adjacent frames as given below:

$$\pi(\mathbf{x} : \mathbf{c}, r) = \begin{cases} 2(1 - \frac{|\mathbf{x}-\mathbf{c}|}{r})^2, & \textit{for} \quad \frac{r}{2} < |\mathbf{x} - \mathbf{c}| < r, \\ 1 - 2(\frac{|\mathbf{x}-\mathbf{c}|}{r})^2, & \text{for} \quad 0 < |\mathbf{x} - \mathbf{c}| < \frac{r}{2}, \\ 0, & \qquad \textit{otherwise}, \end{cases} \tag{7.6}$$

where X is the vector of values of a feature in n adjacent frames, c is the mean vector of x's for all patterns, and r is the radius of the n-dimensional π function. The radius of the n-dimensional π function is obtained from the radii, r_i, of one-dimensional π functions of feature in individual frames.

$$r = \sqrt{\sum_{1=1}^{n} r_i{}^2} \tag{7.7}$$

A two-dimensional π function was used in our studies. We present the effects of the methods of fuzzification on the classification of SCV transitions in the next section.

7.9 Performance for Different Fuzzification Methods

For each SCV transition, a fixed 40 ms portion around the vowel onset point was considered. Formants were extracted using group delay technique [86]. The formant contours were hand edited and smoothed to remove spurious peaks. From the resulting smooth contours the first three formants were obtained for each of the 10 frames at a frame rate of 3.2 ms. The formant data is fuzzified using methods discussed in the previous two sections. Thus for each SCV transition, a 90-dimensional vector of membership values is generated. This representation is used as input to a a.multilayer perceptron network with two hidden layers. The desired output specified during training is also fuzzified. The multispeaker data sets for SCV transitions considered in the studies on parametric representations were used for training and testing. The classification performance on the test data for different methods of fuzzification is given in Table-7.3.

The results show that fuzzification of input and output data improves the classification accuracy. In particular, fuzzification of input data taking into account the fact that the formant data is for a sequence of frames, improves the performance significantly. In these studies only a simple method was used to implement the dependence of fuzziness on the sequence. But a more sophisticated data dependent approach for determining the fuzzy membership values may improve the performance still further.

Table 7.3: Comparison of performance for different fuzzification methods in classification of SCV transitions. The performance is given for two cases of deciding the class: (1) Correct class is the class with the highest output and (2) Correct class is amongst the classes with the highest and the second highest outputs.

Fuzzification Method	Case-1	Case_2
Non-fuzzy inputs	29.5	46.3
Fuzzification of individual frames	62.9	82.1
Fuzzification by variability reduction	70.2	84.8
Fuzzification using 2-dimensional π function	73.5	85.4

7.10 Conclusions

A summary of major results of studies carried out on classification of SCV transitions is given in Table-7.4.

Table 7.4: Summary of the results of studies on parametric representations for classification of SCV transitions.

1. Classification of SCV transitions accurately is important for correct classification of the place of articulation of stop consonants.

2. The performance of parametric representations based on fixed frame size analysis and pitch region analysis has been compared. The formants and spectral coefficient representations based on pitch region analysis gave a better performance.

3. The pitch region analysis based representations gave a better performance for male speaker data and a poorer performance for female speaker data.

4. Different methods for fuzzification of formant features have been explored for classification of SCV transitions.

In this chapter, we have explored methods for suitable representation of SCV transitions to improve the performance for classification of the place of articulation of stop consonants. The performance of the pitch region analysis based parametric representations have been compared with the fixed frame size analysis based representations. The pitch region analysis based methods gave a better performance for the male speaker data. Suitable techniques for extraction of parameters from short duration segments have to be explored to improve the performance for the female speaker data. In order to account for the variability in the characteristics of SCV transitions, methods for fuzzification of formant features have been explored. The results have indicated that fuzzification of formant trajectories gives an improved accuracy in classification.

Here we have explored suitable representations for the transition regions of SCV segments. Methods for using the information in suitable representations for different production events to perform classification have to be explored in order to obtain an improved performance for recognition of SCV segments.

Chapter 8

SUMMARY AND CONCLUSIONS

Approaches for vocabulary independent continuous speech recognition systems are based on models for classification of subword units. Humans recognize speech by discrimination of sounds and by using different types of knowledge. Acoustic-phonetic knowledge is used to resolve ambiguities at the level of subword units. Lexical, syntactic and semantic knowledge sources are used to resolve ambiguities at word and sentence levels. We believe that the success in exploring methods for improving the performance of recognition systems lies in appropriate usage of these knowledge sources even when the best available representations and discrimination models are used for recognition. In this thesis we have demonstrated the potential of this approach by developing a Constraint Satisfaction Model (CSM) for classification of Stop Consonant-Vowel (SCV) utterances. The acoustic-phonetic knowledge of the SCV classes has been incorporated in the form of constraints in this model, and these constraints do provide improved classification over conventional classifiers. The constraints have been used to enhance even the weak evidences available in the outputs of the neural networks (subnets) trained for subgroups of classes and combine multiple evidences available in the outputs of the subncts based on different criteria for grouping. Further improvement in the performance of the CSM requires better methods for deriving constraints based on the acoustic-phonetic knowledge and experimental evidence. Though the CSM has been developed for the classification of isolated utterances of SCVs, the approach can be extended for the classification of SCV segments in continuous speech.

Approaches for spotting subword units have been based on scanning the speech signal continuously using the classifiers for the units. We have demonstrated that the number of false alarms in spotting the SCV units can be reduced significantly by limiting the scanning to the segments around Vowel Onset Points (VOPs). The computational complexity in spotting is also reduced by limiting the scanning. Our approach for spotting can be extended to subword units of all CVs.

The performance of our approaches for classification and spotting of subword units can be improved by using suitable representations for the units and by using models with better discrimination capability. We present some additional research issues in exploring methods for improving the performance.

Exploration of suitable parametric representations can be carried out in the following two ways: (1)Signal dependent analysis and (2) Class dependent analysis. Signal dependent analysis is based on identification of suitable speech parameters for the regions of different significant events in the production of subword units. In the case of SCVs, it involves identification of suitable parameters for the closure, burst, aspiration, transition and vowel regions. A single parametric representation may not be suitable for all the regions. It is necessary to evolve representations that can capture the discriminatory clues in each of the regions. Multiple parametric representations are to be extracted from speech signal of a subword unit segment and given as input to the classifiers.

In the approach based on class dependent analysis, suitable parametric representations for subgroups of classes can be identified. The patterns derived from the speech signal can be based on different parametric representations for different subgroups. In modular neural networks, a separate network is trained for each subgroup. Therefore one can use a parametric representation that is suitable for discrimination among the classes in a subgroup. For example, in grouping based on the manner of articulation

(MOA) of the stop consonants, all the SCV classes in a subgroup have the same MOA. Therefore one can use parametric representations that focus on capturing the clues for discrimination of the place of articulation and the vowel characteristics in deriving the patterns. The results of our studies on classification have shown that the performance for aspirated **SCVs** is poor. It is possible to improve the performance by using a suitable parametric representation for them. The class dependent analysis based approach provides the scope for continuously evaluating the performance for different classes and refining the representations to improve the performance.

In addition to identifying suitable parametric representations, it is important to explore methods for extracting features that can absorb variations in the representations due to the contextual effects and different speakers. Fuzzy logic based methods can be explored for deriving features from the parametric representations in such a way that it is easier for classifiers to perform discrimination in the feature domain.

The performance of classifiers for large number of similar classes as in recognition of **subword** units is dependent on their ability to form complex decision surfaces in the input feature space. The shapes of the decision surfaces vary for different neural network models and architectures. They are also dependent on the methods used for training. As it is difficult to visualize the shapes of the decision surfaces in a large dimensional feature space, it is important to explore analytical methods that can give an insight into the discrimination capability of classifiers.

Modular network architectures have to be necessarily considered for large number of classes. Different grouping criteria lead to different subgroups of classes and hence the shape of decision surface for a class varies for each grouping. In this context, the interpretation of neural networks trained for subgroups of classes as nonlinear filters can be used to analyse the performance for each class. The distribution of the outputs of filters for a class can be analysed to identify the classes that are close to it. The results of this analysis can be used to continuously refine the models and improve the performance for each class.

In our studies on classification of SCVs, fixed duration patterns derived from varying duration segments and utterances have been used for training and testing the classifiers. It was assumed that these patterns have all the necessary information for classification. The performance is dependent on the robustness of the method used for the detection of VOPs in SCV segments and the durations of the portions before and after VOP used for deriving the patterns. Loss of crucial discriminatory information in the process of deriving the patterns can lead to errors in classification. Therefore it is important to explore temporal processing neural network models [40] to handle varying duration patterns. Another advantage of using temporal processing models is that it is possible to train them with isolated utterance data of subword units and use them for classification or spotting of segments of units with different durations in continuous speech.

In conclusion, the following is a list of topics that need further study:

1. Refinement of the constraint satisfaction model to incorporate all the available evidence for solving the recognition problem.

2. Development of robust spotting techniques to take care of the variability in continuous speech.

3. Organization of modular networks to capture the discriminability among the classes.

4. Representation of the speech information taking into account the deterministic, stochastic, fuzzy and temporal nature of the features in the input data.

Appendix A

Algorithm for Extraction of Weighted Cepstral Coefficients

In this appendix, we present the algorithm used for extraction of weighted cepstral coefficient representation from speech signal. This algorithm is taken from [14](pages 112-117).

The digitized speech signal, $s(n)$, is preemphasized by implementing the following difference equation:

$$\tilde{s}(n) = s(n) - 0.95 * s(n-1) \tag{A.1}$$

Let N be the frame size and M be the separation between adjacent frames specified in number of speech signal samples. Then the l**th** frame of speech is denoted by

$$x_l(n) = \tilde{s}(Ml + \mathrm{n}), \qquad n = 0, 1, ..., N-1, l = 0, 1, ..., \mathrm{L} - 1. \tag{A.2}$$

where L is the number of frames in the entire speech signal. Each frame is windowed using a Hamming window as given below.

$$\tilde{x}_l(n) = x_l(n)w(n), \qquad 0 \leq n \leq N-1. \tag{A.3}$$

$$w(n) = 0.54 - 0.46cos(\frac{2n\pi}{N-1}), \qquad 0 \leq n \leq N-1. \tag{A.4}$$

Each frame of windowed signal is autocorrelated as follows:

$$\tilde{r}_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m), \qquad m = 0, 1, ..., p, \tag{A.5}$$

where p is the order of the linear prediction analysis. Linear prediction coefficients are derived from autowrrelation coefficients using Durbin's method given below. The subscript 1 is omitted for convenience.

$$E^{(0)} = r(0) \tag{A.6}$$

$$k_i = \frac{r(i) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} r(|i-j|)}{E^{(i-1)}}, \qquad 1 \le i \le p. \tag{A.7}$$

$$\alpha_i^{(i)} = k_i \tag{A.8}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \tag{A.9}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \tag{A.10}$$

The above equations are solved recursively for i = 1,2,...,p, and the final solution gives the linear prediction coefficients, a_m, as follows:

$$a_m = \alpha_m^{(p)}, \qquad 1 \le m \le p \tag{A.11}$$

The cepstral coefficients, c_m, are derived from linear prediction coefficients by recursion of the following equations:

$$c_0 = ln\sigma^2 \tag{A.12}$$

$$c_m = a_m + \sum_{k=1}^{m-1} (\frac{k}{m}) c_k a_{m-k}, \qquad 1 \le m \le p \tag{A.13}$$

$$c_m = \sum_{k=1}^{m-1} (\frac{k}{m}) c_k a_{m-k}, \qquad p \le m \le Q \tag{A.14}$$

where σ^2 is the gain term in linear prediction analysis and Q is the number of cepstral coefficients. The cepstral coefficients are weighted using a bandpass filter in the cepstral domain as given below to obtain weighted cepstral coefficients, $\hat{c}_m$.

$$c_m = w_m c_m \tag{A.15}$$

where

$$w_m = 1 + \frac{Q}{2} sin(\frac{m\pi}{Q}), \qquad 1 \le m \le Q \tag{A.16}$$

Appendix B

Learning Algorithm for Time Delay Neural Network

The instantaneous sum of squared errors, E, in the standard backpropagation algorithm for multilayer perceptron is defined **as:**

$$E = \frac{1}{2}\sum_{k=1}^{M} e_k^2 \tag{B.1}$$

where e_k is the error at the output of the kth neuron in the output layer and M is the number of nodes in the output layer. The error e_k is defined *as:*

$$e_k = (d_k - y_k) \tag{B.2}$$

where d_k is the desired target value of the kth neuron in the output layer and y_k is the actual output of the kth neuron.

For TDNN, the output value y_k for the kth output node is defined as the average of the outputs of the replicas of the Eth output node and therefore is given as:

$$y_k = \frac{1}{N}\sum_{l=1}^{N} y_{kl} \tag{B.3}$$

where y_{kl} **is** the actual output of the lth replica of the kth output node, and N is the number of replicas for each output node. The error for all the N replicas of the kth node is defined as:

$$e_k = e_{kl} = d_k - y_k = d_k - \frac{1}{N}\sum_{l=1}^{N} y_{kl} \tag{B.4}$$

Therefore, the instantaneous sum of squared errors, E, for TDNN is derived as:

$$E = \frac{1}{2}\sum_{k=1}^{M}(d_k - \frac{1}{N}\sum_{l=1}^{N} y_{kl})^2 \tag{B.5}$$

The correction Δw_{ijkl} applied to the weight w_{ijkl} of the connection between the jth replica of the ith node in the hidden layer and the lth replica of the kth node in the output layer is defined as:

$$\Delta w_{ijkl} = -\eta \frac{\partial E}{\partial w_{ijkl}} \tag{B.6}$$

where the partial derivative of E with respect to w_{ijkl} is the instant gradient and is expressed by the chain rule as follows:

$$\frac{\partial E}{\partial w_{ijkl}} = \frac{\partial E}{\partial e_{kl}}\frac{\partial e_{kl}}{\partial y_{kl}}\frac{\partial y_{kl}}{\partial v_{kl}}\frac{\partial v_{kl}}{\partial w_{ijkl}} \tag{B.7}$$

where v_{kl} is the net internal activity level of the lth replica of the kth node. $v_{kl}$is given by

$$v_{kl} = \sum_{i=1}^{L}\sum_{j\in S_l} w_{ijkl}y_{ij} \tag{B.8}$$

where L is the number of nodes in the hidden layer, S_l is the set of hidden layer columns in the receptive field of the lth replica, and y_{ij} is the output of the jth replica of the ith hidden node.

The expressions for the partial derivatives in (7) are derived as below:

$$\frac{\partial E}{\partial e_{kl}} = e_{kl} \tag{B.9}$$

$$\frac{\partial e_{kl}}{\partial y_{kl}} = -\frac{1}{N} \tag{B.lO}$$

$$\frac{\partial y_{kl}}{\partial v_{kl}} = \Phi'(v_{kl}) \tag{B.11}$$

$$\frac{\partial v_{kl}}{\partial w_{ijkl}} = y_{ij} \tag{B.12}$$

Here $\Phi(x)$ is the activation function of the neurons and $\Phi'(x)$ is the partial derivative of the activation function with its argument x.

Substituting these expressions in (7) and (6), the correction in the weight w_{ijkl} is expressed as:

$$\Delta w_{ijkl} = \frac{1}{N}\eta e_{kl}\Phi'(v_{kl})y_{ij} \tag{B.13}$$

The local gradient δ_{kl} is defined as

$$\delta_{kl} = \frac{1}{N}e_{kl}\Phi'(v_{kl}) \tag{B.14}$$

Then Δw_{ijkl} can be expressed as

$$\Delta w_{ijkl} = \eta\delta_{kl}y_{ij} \tag{B.15}$$

A similar procedure is used to derive the expression for the correction in the weight w_{pqij} of the connection between the qth unit of the pth frame in the input layer and the jth replica of ith node in the hidden layer.

$$\Delta w_{pqij} = \eta\delta_{ij}y_{pq} \tag{B.16}$$

where δ_{ij} is the local gradient for the jth replica of the ith hidden node and y_{pq} is qth input value of the pth frame. δ_{ij} is defined as:

$$\begin{aligned}\delta_{ij} &= -\frac{\partial E}{\partial y_{ij}}\frac{\partial y_{ij}}{\partial v_{ij}} && \text{(B.17)}\\ &= -\frac{\partial E}{\partial y_{ij}}\Phi'(v_{ij}) && \text{(B.18)}\end{aligned}$$

$$\frac{\partial E}{\partial y_{ij}} = \sum_{k=1}^{M}\sum_{l\in S_j} e_{kl}\frac{\partial e_{kl}}{\partial y_{ij}} \tag{B.19}$$

where S_j is the set of columns in the output layer to which there is a connection for the jth replica of the ith hidden node. From equations (10), (11) and (8), we get

$$\frac{\partial e_{kl}}{\partial y_{ij}} = \frac{\partial e_{kl}}{\partial y_{kl}}\frac{\partial y_{kl}}{\partial v_{kl}}\frac{\partial v_{kl}}{\partial y_{ij}} \tag{B.20}$$

$$= -\frac{1}{N}\Phi'(v_{kl})w_{ijkl} \tag{B.21}$$

Substituting this in (19) and and then using (14), we get

$$\frac{\partial E}{\partial y_{ij}} = -\sum_{k=1}^{M}\sum_{l\in S_j} e_{kl}\frac{1}{N}\Phi'(v_{kl})w_{ijkl} \tag{B.22}$$

$$= -\sum_{k=1}^{M}\sum_{l\in S_j} \delta_{kl}w_{ijkl} \tag{B.23}$$

Substituting this expression in (18), δ_{ij} is derived as

$$\delta_{ij} = \Phi'(v_{ij})\sum_{k=1}^{M}\sum_{l\in S_j} \delta_{kl}w_{ijkl} \tag{B.24}$$

Finally, the expression for Δw_{pqij} is obtained as

$$\Delta w_{pqij} = \eta\Phi'(v_{ij})y_{pq}\sum_{k=1}^{M}\sum_{l\in S_j} \delta_{kl}w_{ijkl} \tag{B.25}$$

Appendix C

Training and Recognition Algorithms for Discrete Hidden Markov Models

In this appendix, we present the training and recognition algorithms for discrete hidden Markov models used in our studies on classification of SCVs. The algorithms are taken from [14](pages 329-370).

A Discrete Hidden Markov Model (DHMM) is characterized by the following:

1. N, the number of states in the model.

2. M, the number of distinct observation symbols per state. In our studies, M corresponds to the size of the codebook built from vector quantization of weighted cepstral coefficient vectors. The individual symbols are denoted as V = $\{v_1, v_2, \ldots, v_M\}$ and in our case v_i's refer to the codebook indices.

3. The state-transition probability distribution, A = $\{a_{ij}\}$ where a_{ij} is the probability of making a transition from state i to state j at time t and is given by

   $$a_{ij} = P[q_{t+1} = j | q_t = i], \qquad 1 \leq i, j \leq N. \tag{C.1}$$

   Here q_t refers to the state of the model at time t.

4. The observation symbol probability distribution, B = $\{b_j(k)\}$, in which

   $$b_j(k) = P[o_t = v_k | q_t = \mathrm{j}], \qquad 1 \leq \mathrm{k} \leq \mathrm{M}, \tag{C.2}$$

defines the symbol distribution in state j, j = 1,2,...,N. Here o_t refers the symbol output by the model at time t.

5. The initial state distribution, $\pi = \{\pi_i\}$ in which

$$\pi_i = P[q_1 = i], \qquad 1 \leq i \leq N. \tag{C.3}$$

A DHMM A is completely specified by the model parameters N and M, and the probability measures A, B and π. For classification of SCVs, we have used models with 5 states (i.e., N = 5) and a codebook size of 256 (i.e., M = 256). The probability measures are estimated from training data using the Baum-Welch method given below.

C.1 Training Algorithm for DHMM

The DHMM model parameters are estimated from the patterns in training data set using the algorithm given below.

C.1.1 The Forward Procedure

The forward variable $\alpha_t(2)$ is defined as

$$\alpha_t(i) = P(o_1 o_2 ... o_t, q_t = i|\lambda) \tag{C.4}$$

that is, the probability of the partial observation sequence, $o_1 o_2 ... o_t$, and state i at time t, given the model λ. The forward variable $\alpha_t(i)$ is computed inductively, as follows:

1. Initialization

$$\alpha_1(i) = \pi_i b_i(o_1), \qquad 1 \leq i \leq N. \tag{C.5}$$

2. Induction

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i)a_{ij}]b_j(o_{t+1}), \qquad 1 \leq j \leq N, 1 \leq t \leq T-1. \tag{C.6}$$

Here T refers to the length of the complete observation sequence O. In our case T corresponds to the number of frames in the speech signal.

3. Termination

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{C.7}$$

C.1.2 The Backward Procedure

The backward variable $\beta_t(i)$ is defined as

$$\beta_t(i) = P(o_{t+1}o_{t+2}...o_T|q_t = i, \lambda) \tag{C.8}$$

that is, the probability of the partial observation sequence from $t+1$ to the end, given state i at time t and the model λ. The backward variable $\beta_t(i)$ is computed inductively, as follows:

1. Initialization

$$\beta_T(i) = 1, \qquad 1 \leq i \leq N. \tag{C.9}$$

2. Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}b_j(o_{t+1})\beta_{t+1}(j), \qquad 1 \leq i \leq N, t = T-1, T-2, ..., 1. \tag{C.10}$$

C.1.3 Parameter Estimation

Define $\xi_t(i,j)$ as the probability of being in state i at time t, and state j at time $t+1$, given the model λ and the observation sequence O, and is given as follows:

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j|O, \lambda). \tag{C.11}$$

From the definitions of the forward and backward variables, the following expression is derived for $\xi_t(i,j)$:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)} \tag{C.12}$$

Define $\gamma_t(i)$ as the probability of being in state i at time t, given the entire observation sequence and the model. Then $\gamma_t(i)$ is related to $\xi_t(i,j)$ as follows:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j). \tag{C.13}$$

If we define the current model for DHMM by $\lambda = (A, B, \pi)$, then the reestimated model $\bar{\lambda} = (A, B, \bar{\pi})$ is computed using the following formulae:

$$\bar{\pi}_i = \gamma_1(i) \tag{C.14}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{C.15}$$

$$\bar{b}_j(k) = \frac{\sum_{t=1, s.t. o_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{C.16}$$

Based on the above procedure, $\bar{\lambda}$ is iteratively used in place of λ and the reestimation is calculated repeatedly until the probability of O being observed from the model reaches a limiting point. The final result of the reestimation procedure a maximum likelihood estimate of the DHMM for the observation sequence O.

The above reestimation procedure is extended for training a DHMM with multiple observation sequences, i.e., multiple number of patterns, as follows:

Let us denote the set of K observation sequences as

$$O = [O^{(1)}, O^{(2)}, ..., O^{(K)}] \tag{C.17}$$

where $O^{(k)} = (o_1^{(k)}, o_2^{(k)}, ..., o_{T_k}^{(k)})$ is the kth observation sequence and T_k is the length of the kth observation sequence. The modified reestimation formulae are as follows:

$$\bar{a}_{ij} = \frac{\sum\limits_{k=1}^{K} \frac{1}{P_k} \sum\limits_{t=1}^{T_k-1} \alpha_t^k(i) a_{ij} b_j(o_{t+1}^{(k)}) \beta_{t+1}^k(j)}{\sum\limits_{k=1}^{K} \frac{1}{P_k} \sum\limits_{t=1}^{T_k-1} \alpha_t^k(i) \beta_t^k(i)} \tag{C.18}$$

$$\bar{b}_j(l) = \frac{\sum\limits_{k=1}^{K} \frac{1}{P_k} \sum\limits_{t=1, s.t. o_t = v_l}^{T_k-1} \alpha_t^k(i) \beta_t^k(i)}{\sum\limits_{k=1}^{K} \frac{1}{P_k} \sum\limits_{t=1}^{T_k-1} \alpha_t^k(i) \beta_t^k(i)} \tag{C.19}$$

The above reestimation procedure is used for training a DHMM model for each of the SCV classes using patterns in the training data set.

C.2 Recognition Algorithm for DHMM

To use the DHMM specified by λ for recognition of the observation sequence O of a pattern, the probability $P[O|\lambda]$ is computed using the forward procedure. Given an SCV pattern, these probabilities are computed for the models of all the classes and the class with the highest probability is assigned to the given pattern.

REFERENCES

[1] J.B.Allen, "How do humans process and recognize speech?," IEEE Transactions on Speech and Audio Processing, vol. 2, pp. 567–577, October 1994.

[2] P.Eswar, A Rule-based Approach for Spotting Characters from Continuous Speech in Indian Languages. PhD thesis, I.I.T., Madras, July 1990.

[3] F.Jelinek, "Continuous speech recognition by statistical methods," Proceedings of the IEEE, vol. 64, pp. 532–555, April 1976.

[4] J.Picone, "Continuous speech recognition using hidden Markov modelling," IEEE ASSP Magazine, vol. 7, pp. 26–41, July 1990.

[5] B.Yegnanarayana, "Artificial neural networks for pattern recognition," Sadhana, vol. 19, pp. 189–238, April 1994.

[6] R.P.Lippmann, "Review of neural networks for speech recognition," Neural Computation, vol. 1, pp. 1–38, 1989.

[7] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.J.Lang, "Phoneme recognition using time-delay neural networks," IEEE Transactions on Acoustics Speech and Signal Processing, vol. 37, pp. 328–339, March 1989.

[8] N.Morgan and H.Bourlard, "Continuous speech recognition - An introduction to the hybrid HMM/connectionist approach," IEEE Signal Processing Magazine, vol. 12, pp. 24–42, May 1995.

[9] R.Sundar, Signal Processing for Recognition of Isolated Utterances of Speech Units. PhD thesis, I.I.T., Madras, January 1994.

[10] S.E.G.Ohman, "Coarticulation in VCV utterances: Spectrographic measurements," Journal of Acoustic Society of America, vol. 31, pp. 151–168, 1966.

[11] D.Kewley-Port, "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," Journal of Acoustic Society of America, vol. 72, pp. 379–389, 1982.

[12] K.N.Stevens, "Models for production and acoustics of stop consonants," Speech Communication, vol. 13, pp. 367–375, December 1993.

[13] P.van Alphen, *HMM-based* Continuous Speech *Recognit*ion: Systematic Evaluation of Various System Components. University of Amsterdam, 1992.

[14] L.R.Rabiner and B.H.Juang, Fundamentals of Speech Recognition. Prentice-Hall, 1993.

[15] K.F.Lee, Automatic Speech Recognition: The Development of the SPHINX System. Kluwer, 1989.

[16] K.F.Lee and H.W.Hon, "Speaker-independent phone recognition using hidden Markov models," Tech. Rep. CMU-CS-88-121, Carnegie-Mellon University, March 1988.

[17] K.F.Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," IEEE *Transactions* on Acoustics Speech and Signal Processing, vol. 38, pp. 599–609, April 1990.

[18] C.H.Lee, L.R.Rabiner, R.Pieraccini, and J.G.Wilpon, 'Acoustic modeling for large vocabulary speech recognition," Computer Speech and Language, pp. 137–165, January 1990.

[19] C.H.Lee, L.R.Rabiner, and R.Pieraccini, "Speaker independent continuous speech recognition using continuous density hidden Markov models," in Speech Recognition and Understanding: Recent Advances, *Trends* and Applications (P.Laface and R. Mori, eds.), pp. 135–163, Springer-Verlag, 1992.

[20] L.Deng, M.Lennig, and P.Mermelstein, 'Modelling microsegments of stop consonants in a hidden Markov model based word recognizer," Journal of Acoustic Society of America, vol. 87, pp. 2738–2747, June 1990.

[21] L.Deng and K.Erler, "Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech unit," Journal of Acoustic Society of America, vol. 92, pp. 3058–3067, December 1992.

[22] T.Waardenburg, J. Preez, and M.W.Coetzer, "The automatic recognition of stop consonants using hidden Markov models," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, vol. 1, pp. 585–588, 1992.

[23] H.Bourlard, N.Morgan, and S.Renals, "Neural nets and hidden markov models: Review and generalizations," Speech Communication, vol. 11, pp. 237–246, June 1992.

[24] L.R.Bahl, P.F.Brown, P. Souza, and L.R.Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, pp. 49–52, 1986.

[25] K.J.Lang and G.E.Hinton, "A time-delay neural network architecture for speech recognition," Tech. Rep. CMU-CS-88-152, Carnegie-Mellon University, December 1988.

[26] T.D.Harrison and F.Fallside, "A connectionist model for phoneme recognition in continuous speech," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, pp. 417–420, 1991.

[27] S.Renals and R.Rower, "Learning phoneme recognition using neural networks," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, pp. 413–416, 1989.

[28] K.K.Paliwal, "Neural net classifiers for robust speech recognition under noisy environment," in Proceeding of the International Conference on *Acoustics* Speech and Signal Processing, pp. 429-432, 1990.

[29] M.Kokkonen and K.Torkkola, "Using self-organizing maps and multilayered feed-forward nets to obtain phonemic transcriptions of spoken utterances," in Proceedings of European Conference on Speech Processing and Technology, vol. 2, pp. 561–564, 1989.

[30] T.Kohonen, "Speech recognition based on topology preserving neural maps," in Neural Computing Architectures: The Design of Brain-Like Machines (I.Alexander, ed.), pp. 26-40, North Oxford Academic, 1989.

[31] R.Cole, M.Fanty, Y.Muthuswamy, and M.Gopalakrishnan, "Speaker-independent recognition of spoken English letters," in Proceeding of the International Joint Conference on Neural Networks, 1990.

[32] P.F.Brown, The Acoustic-Modelling Problem in Automatic Speech Recognition. PhD thesis, Carnegie Mellon University, 1987.

[33] K.-Y.Su and C.H.Lee, "Speech recognition using weighted HMM and subspace projection approaches," IEEE Transactions on Speech and Audio Processing, vol. 2, pp. 69–79, January 1994.

[34] L.A.Zadeh, "Fuzzy logic," IEEE Computer, vol. 21, pp. 83–93, April 1988.

[35] S.K.Pal and D. Majumder, Fuzzy Mathematical Approach to Pattern Recognition. Wiley, 1986.

[36] P.K.Simpson, "Fuzzy min-max neural networks - Part 1: Classification," IEEE Transactions on Neural Networks, vol. 3, pp. 776–786, September 1992.

[37] W.Pedrycz, "Fuzzy neural networks with reference neurons as pattern classifiers," *ILEE* Transactions on Neural Networks, vol. 3, pp. 770–775, September 1992.

[38] S.K.Pal and S.Mitra, "M'ultilayer perceptron and fuzzy sets and classification," IEEE Transactions on Neural Networks, vol. 3, pp. 683–697, September 1992.

[39] Y.-H.Kuo, C.-I.Kao, and J.-J.Chen, "A fuzzy neural network model and its hardware implementation," IEEE Transactions on Fuzzy Systems, pp. 171–183, August 1993.

[40] S.Haykin, Neural Networks: A Comprehensive Foundation. Macmillan, 1994.

[41] A.Waibel, H.Sawai, and K.Shikano, 'Modularity and scaling in large phonemic neural networks," IEEE Transactions on Acoustics Speech and Signal Processing, vol. 37, pp. 1888–1898, December 1989.

[42] H.Sawai, A.Waibel, M.Miyatake, and K.Shikano, "Spotting Japanese CV syllables and phonemes using time-delay neural networks," in Proceeding of the International Conference on *Acoustics* Speech and Signal Processing, pp. 25–28, May 1989.

[43] N.Morgan and H.A.Bourlard, "Neural networks for statistical recognition of continuous speech," Proceedings of the IEEE, vol. 83, pp. 742–770, May 1995.

[44] J.G.Wilpon, L.R.Rabiner, C.H.Lee, and E.R.Goldman, 'Automatic recognition of keywords in unconstrained speech using hidden Markov models," IEEE Transactions on Acoustics Speech and Signal Processing, vol. 38, pp. 1870–1878, November 1990.

[45] R.C.Rose and D.B.Paul, "A hidden Markov model based keyword recognition system," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, pp. 129–132, 1990.

[46] J.R.Rohlicek, W.Russel, S.Roukos, and H.Gish, "Continuous hidden Markov modelling for speaker-independent word spotting," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, pp. 627–630, 1989.

[47] R.C.Rose, "Discriminant wordspotting techniques for rejecting non-vocabulary utterances in unconstrained speech," in Proceeding of the *International* Conference on Acoustics Speech and Signal Processing, vol. II, pp. 105–108, 1992.

[48] T.Komori and S.Katagiri, "A new learning algorithm for minimizing spotting errors," in Proceedings of IEEE Workshop on Neural Networks and Signal Processing, pp. 333–342, 1993.

[49] T.M.English and L.C.Boggess, "Back-propagation training of a neural network for word spotting," in Proceeding of the .International Conference on Acoustics Speech and Signal Processing, vol. I, pp. 357–360, 1992.

[50] K.P.Li and J.A.Naylor, "A whole word recurrent neural network for keyword spotting," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, vol. II, pp. 81–84, 1992.

[51] S.V.Kosonocky and R.J.Mammone, 'A continuous density neural tree network word spotting system," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, vol. I, pp. 305–308, 1995.

[52] D.P.Morgan, C.L.Scofield, and J.E.Adcock, "Multiple neural network topologies applied to keyword spotting," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, vol. I, pp. 313–316, 1991.

[53] R.DeMori, M.J.Palakal, and P.Cosi, Perceptual Models for Automatic Speech Recognition Systems, vol. 31, pp. 99–173. Academic Press, 1990.

[54] G.Fant, Speech Sounds and Features. The MIT Press, 1973.

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

[56] J.R.Deller, J.G.Proakis, and J.H.L.Hansen, Discrete Time Processing of Speech Signals. Mac Millan, 1993.

[57] J.Harrington, "Acoustic cues for automatic recognition of English consonants," in Aspects of Speech Technology (M.A.Jack and J.Laver, eds.), Edinburgh University Press, 1988.

[58] M.F.Dorman, M.S.Kennedy, and L.J.Raphel, "Stop consonant recognition: release bursts and formant transitions as functionally equivalent context-dependent cues," Perception and *Psychoacoustics*, vol. 22, pp. 109–122, 1977.

[59] R.Smits, Detailed versus Gross *Spectro-*Temporal Cues for the Perception of Stop Consonants. PhD thesis, Technical University of Eindhoven, 1995.

[60] D.J.Sharf and T.Hemeyer, "Identification of place of consonant articulation from vowel formant transitions," Journal of Acoustic Society of America, no. 51, pp. 652–658, 1972.

[61] K.Etemad, "Phoneme recognition based on multi-resolution and non- causal context," in *Proceedings* of IEEE Workshop on Neural *Networks* and Signal Processing, pp. 343–352, 1993.

[62] M.E.H.Schouten and L.C.W.Pols, "Vowel segments in consonantal contexts: A spectral study of coarticulation - Part I," *Jl.* of Phonetics, vol. 7, no. 1, pp. 1–23, 1979.

[63] P.C.Delattre, A.M.Liberman, and F.S.Cooper, "Acoustic loci and transitional cues for stop consonants," Journal of Acoustic Society of America, vol. 27, pp. 769–773, 1955.

[64] L.Deng, P.Kenny, M.Lennig, and P.Mermelstein, "Modelling acoustic transitions in speech by state-interpolation hidden Markov models," IEEE Transactions on Signal Processing, vol. 40, pp. 265–272, February 1992.

[65] H.M.Sussman, H.A.McCaffrey, and S.A.Mathews, "An investigation of locus equations as a source of relational invariance for stop place categorization," Journal of Acoustic Society of America, vol. 90, pp. 1309–1325, 1991.

[66] L.Deng and D.Braam, "Context-dependent Markov model structured by locus equations: Applications to phonetic classification," Journal of Acoustic Society of America, vol. 96, pp. 2008–2025, October 1994.

[67] L.Deng, "A statistical model for formant-transition microsegments of speech incorporating locus equations," Signal Processing, vol. 37, pp. 121–128, May 1994.

[68] K.S.Nathan, Y.T.Lee, and H.F.Silverman, "A time-varying analysis method for rapidly changing signals," IEEE Transactions on Signal Processing, vol. 39, pp. 815–824, 1991.

[69] K.S.Nathan and H.F.Silverman, "Time-varying feature selection and classification of unvoiced stop consonants," IEEE Transactions on Speech and Audio Processing, vol. 2, pp. 395–405, July 1994.

[70] C.Chandra Sekhar and B.Yegnanarayana, "Classification of CV transitions in continuous speech using neural network models," in Proceedings of International Symposium on Speech, Image Processing and Neural Networks, pp. 97–100, 1994.

[71] H.Kasuya and H.Wakita, "An approach to segmenting speech into vowel- and nonvowel-like intervals," IEEE Transactions on Acoustics Speech and Signal Processing, vol. 27, pp. 319–327, 1979.

[72] D.J.Hermes, "Vowel-onset detection," Journal of *Acoustic* Society of America, vol. 87, pp. 866–873, February 1990.

[73] A.Bendiksen and K.Steiglitz, "Neural networks for voiced/unvoiced speech classification," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, pp. 521–524, April 1990.

[74] Y.Qi and B.R.Hunt, "Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier," IEEE Transactions on Speech and Audio Processing, vol. 1, pp. 250–255, April 1993.

[75] B.Yegnanarayana and R.Teunen, "Prosodic manipulation of speech using knowledge of instants of significant excitation," Tech. Rep. 1029, Institution of Perceptional Research, IPO, Eindhoven, The Netherlands, 1994.

[76] R.Smits and B.Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," IEEE Transactions on Speech and Audio Processing, vol. 3, pp. 325–333, September 1995.

[77] B.Yegnanarayana and D. Reddy, "A distance measure based on the derivative of linear prediction phase spectrum," in Proceeding of the International Conference on Acoustics Speech and Signal Processing, pp. 744–747, 1979.

[78] D.P.Morgan and C.L.Scofield, Neural Networks and Speech Processing. Kluwer Academic Publishers, 1991.

[79] I.Rogina and A.Waibel, "Learning state-dependent stream weights for multi-codebook HMM speech recognition systems," in Proceeding *of* the International Conference on Acoustics Speech and Signal Processing, vol. I, pp. 217–220, April 1994.

[80] **D.E.Rumelhart, P.Smolensky, J.L.McClelland,** and **G.E.Hinton,** 'Schemata and sequential thought processes in PDP models," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. **Vol.2.** Psychological and Biological Models* (**J.L.McClelland, D.E.Rumelhart,** and the PDP Research Group, eds.), MIT Press, 1986.

[81] **P.P.Raghu** and **B.Yegnanarayana,** "Texture classification using a probabilistic neural network and constraint satisfaction model," in *Proceeding of the International Conference on Neural Networks,* 1996.

[82] **D.E.Rumelhart, J.L.McClelland,** and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol.1. Foundations.* MIT Press, 1986.

[83] **K.N.Stevens** and **S.E.Blumstein,** "Invariant cues for place of articulation in stop consonants," *Journal of Acoustic Society of America,* vol. 64, pp. 1358–1368, November 1978.

[84] **D.Kewley-Port,** "Time-varying features as correlates of place of articulation in stop consonants," *Journal of Acoustic Society of America,* vol. 73, pp. 322–335, 1983.

[85] **A.K.Datta, N.R.Ganguli,** and **S.Ray,** "Recognition of unaspirated plosives - A statistical approach," *IEEE Transactions on Acoustics Speech and Signal Processing,* vol. 28, pp. 85–91, February 1980.

[86] **H.A.Murthy** and **B.Yegnanarayana,** "Formant extraction from group delay function," *Speech Communication,* vol. 10, pp. 209–221, 1991.

LIST OF PAPERS BASED ON THESIS

Journals:

1. C.Chandra Sekhar and B.Yegnanarayana, "Recognition of Stop-Consonant-Vowel (SCV) segments in continuous speech using neural network models," Special Issue of JIETE on Neural Networks, 1996.(Accepted).

2. C.Chandra Sekhar, Santosh S. Mathews and B.Yegnanarayana, "A modular approach for recognition of isolated Stop Consonant- Vowel (SCV) utterances in Indian languages," Journal of The Acoustical Society of India, vol.XXIII, no.1, pp.28-35, 1995.

Conferences:

1. C.Chandra Sekhar and B.Yegnanarayana, "Neural network models for spotting Stop Consonant-Vowel (SCV) segments in continuous speech," International Conference on Neural Networks, ICNN'96, Washington,D.C., June 1996. (Accepted).

2. C.Chandra Sekhar, K.Sakthivel and B.Yegnanarayana, "Neural network Lased approaches for recognition of Stop Consonant Vowel (SCV) segments in continuous speech," Proc. of National Conference on Neural Networks and Fuzzy Systems, Madras, pp.16-22, Mar. 1995.

3. B.Yegnanarayana, C.Chandra Sekhar and S.R.Prakash, "Fuzzification of formant trajectories for classification of CV utterances using neural network models," Proc. of 1994 IEEE Workshop on Neural Networks for Signal Processing, Greece, Sep. 1994.

4. B.Yegnanarayana, S.R.Prakash and C.Chandra Sekhar, "Recognition of CV segments using fuzzy neural networks," Proc. of World Congress on Neural Networks, San Diego, USA, June 1994.

5. C.Chandra Sekhar and B.Yegnanarayana, "Classification of CV transitions in continuous speech using neural network models," Proc. of International Symposium on Speech, Image Processing and Neural Networks, Hong Kong, pp.97-100, Apr. 1994.

NEURAL NETWORK MODELS FOR RECOGNITION OF STOP CONSONANT-VOWEL (SCV) SEGMENTS IN CONTINUOUS SPEECH

A ***THESIS***

submitted for the award of the degree

of

DOCTOR OF PHILOSOPHY

by

C.CHANDRA SEKHAR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY, MADRAS.

APRIL 1996