

COARTICULATION KNOWLEDGE FOR A TEXT-TO-SPEECH SYSTEM FOR AN INDIAN LANGUAGE

A THESIS

submitted for the award of the degree

of

MASTER OF SCIENCE

in

COMPUTER SCIENCE AND ENGINEERING

by

RAMACHANDRAN V.R.

Under the guidance of

Prof. B. YEGNANARAYANA

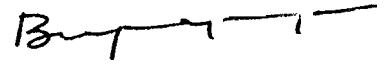


DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
MADRAS - 600 036

MARCH 1993

CERTIFICATE

This is to certify that the thesis entitled "COARTICULATION KNOWLEDGE FOR A TEXT-TO-SPEECH SYSTEM FOR AN INDIAN LANGUAGE" is the bonafide work of Mr. Ramachandran V.R, carried out under my guidance and supervision, in the Department of Computer Science and Engineering, Indian Institute of Technology, Madras, for the award of the degree of Master of Science in Computer Science and Engineering.



(B. Yegnanarayana)

ACKNOWLEDGEMENTS

I am extremely grateful to my guide **Prof.B.Yegnanarayana** for providing me constant motivation and support in this work. He introduced me to the fascinating field of speech synthesis. I am very fortunate for having numerous technical discussions with him from which I benefited enormously. Also he spent lots of his valuable time sitting with me to discuss the implementation details. I am highly indebted to him for supporting me through a **DoE** project throughout this work.

This work was carried out while developing a text-to-speech system for Hindi under the **DoE** project "Interactive Voice Input/Output System for a Computer". I am highly grateful to the Department of Electronics for their support.

This work owes a lot to my friend and colleague **Mr.S.Rajedran** without whom this would not have come to the present stage. He had been a rich source of linguistic knowledge which was very essential for this work. He helped me in the process of speech data collection, analysis, tuning etc. He also proof read this thesis several times patiently. His help is gratefully acknowledged.

I thank **Mr.S.R.Rajesh Kumar** and **Mr.C.Chandra Sekhar** and **Ms.Hema A.Murthy** for their valuable technical help in the early stages of this work.

I thank my friends and co-researchers **Mr.A.S.Madhukumar**, **Mr.R.Ramaseshan**, **Mr.A.Ravichandran** and **Mr.A.Arul Valan** for their encouragement and advice in both technical and nontechnical matters.

I thank my friend and colleague **Mr.K.G.Udayakumar** for his technical help in preparing most of the figures in this thesis which saved me from the tedious job of cut and paste and **Mr.M.Narendranath** for helping me in documenting the programs.

I thank my friend **Mr.K.Suku** for his encouragement and moral support.

I gratefully acknowledge the help of my colleagues in Speech and Vision Laboratory and Microprocessor Laboratory and also the faculty members **Mr.G.V.Ramana Rao**, **Mr.R.Sundar** and **Mr.N.Alwar** who assisted me directly and indirectly in this work.

CONTENTS

ABSTRACT	1
1. INTRODUCTION	2
1.1 Coarticulation effect and its importance	
in text-to-speech conversion	2
1.2 Issues in developing a text-to-speech system	3
1.2.1 Choice of the synthesis model	5
1.2.2 Data collection for segmental synthesis	5
1.2.3 Knowledge sources for naturalness	5
1.3 Organization of the thesis	6
2. STUDIES ON COARTICULATION - AN OVERVIEW	8
2.1 Introduction	8
2.2 Evidence for coarticulation	9
2.3 Importance of coarticulation	9
2.4 Types of coarticulation	10
2.4.1 Anticipatory coarticulation or forward coarticulation	10
2.4.2 Carry-over coarticulation	11
2.5 Models of coarticulation	11
2.5.1 Kozhevnikov-Chistovich articulatory syllable	12
2.5.2 Wickelgren's context-sensitive model	13
2.5.3 Feature-based models	13
2.5.4 Hierarchical model	14

2.6 Studies on coarticulation for text-to-speech systems for some major languages	14
2.7 Coarticulation knowledge for a text-to-speech system for Hindi	16
2.7.1 Vowel-to-consonant coarticulation	16
2.7.2 Consonant-to-vowel coarticulation	16
2.7.3 Vowel-to-vowel coarticulation	17
2.7.4 Nasalization	17
2.7.5 Consonant-to-consonant coarticulation	19
2.8 Factors affecting coarticulation	20
2.9 Summary	20
3. ARCHITECTURE OF A TEXT-TO-SPEECH SYSTEM FOR HINDI	21
3.1 Introduction	21
3.2 Speech synthesis model	21
3.3 Various phases in text-to-speech conversion	23
3.3.1 Input	23
3.3.2 Preprocessor	23
3.3.3 Parser	25
3.3.4 Synthesis of speech from parameters of the basic units	26
3.4 Basic: units and their representation	28
3.4.1 Choice of basic unit for synthesis	28
3.4.2 Representation of basic units	30
3.4.3 Construction of the basic unit data base	37

3.5 Summary

4. COARTICULATION KNOWLEDGE FOR THE

TEXT-TO-SPEECH SYSTEM 43

4.1 Introduction 43

4.2 Nature of coarticulation in Hindi 43

4.3 Acquisition and formulation of coarticulation knowledge for a text-to-speech system 44

4.3.1 Identification of the domain of coarticulation 44

4.3.2 Classification of various coarticulation patterns 45

4.3.3 Formulation of coarticulation patterns as rules 47

4.4 Incorporation of the coarticulation knowledge to the text-to-speech system 68

4.4.1 Architecture of the synthesis module 69

4.4.2 Representation and activation of coarticulation knowledge 70

4.5 Summary 73

5. TESTING AND EVALUATION 74

5.1 Introduction 74

5.2 Testing of the basic units 74

5.3 Testing of the coarticulation rules 75

5.3.1 Experimental testing 75

5.3.2 Perceptual testing of the rules 76

5.4 Evaluation at higher levels 80

5.5 Conclusion	81
6. SUMMARY AND CONCLUSIONS	82
Appendix 1. VAX STATION DETAILS	84
Appendix 2. AN EDITOR FOR BASIC UNIT	
REPRESENTATION AND VC RULES	86
Appendix 3. TABLE OF ISCII CODES	90
Appendix 4. LOOK-UP TABLES FOR THE PREPROCESSOR	94
Appendix 5. HINDI CONSONANTS AND VOWELS	101
Appendix 6. ORGANIZATION OF THE TEXT-TO-SPEECH	
CONVERSION SOFTWARE	104
REFERENCES	107
LIST OF FIGURES	111
LIST OF TABLES	113
LIST OF PUBLICATIONS	114

ABSTRACT

Text-to-speech conversion involves synthesizing speech from the text of a language. In order for the synthesized speech to be natural, we have to model adequately various features of the natural speech. The knowledge pertaining to these features depends on the behaviour of speech production mechanism as well as various linguistic factors. One such knowledge source is the coarticulation effect which is responsible for the smooth flowing nature and hence a factor responsible for naturalness of continuous speech. Coarticulation can be defined as the influence on speech units by neighbouring units in continuous speech, which is caused by physiological constraints of the articulators. This thesis addresses some issues in designing and implementing a text-to-speech system for Hindi, an Indian language, with an emphasis on the problem of acquisition and incorporation of the coarticulation knowledge. Coarticulation gives rise to various transition patterns between adjacent speech units. In our approach to text-to-speech conversion, we use a collection of basic speech units corresponding to the characters of Hindi and use the coarticulation knowledge formulated as a set of contextual rules to join these units to produce continuous speech output.

The issues addressed here are concerned with acquisition and incorporation of the coarticulation knowledge. The coarticulation patterns are decided by the nature of speech units involved and also by the constraints due to physiological factors. These patterns are generally speaker independent. They are derived and formulated from examples of various contexts in natural speech. In order to incorporate the coarticulation patterns while synthesizing speech, we need a representation of the speech units which has the flexibility of spectral and other parameter modification. This is accomplished by using a suitable representation and synthesis scheme using various speech parameters.

The main contributions of this work are the following: (i) The coarticulation knowledge is acquired and formulated as a set of rules, (ii) A representation scheme is proposed for the basic speech units which is flexible enough to allow incorporation of the coarticulation knowledge, (iii) Collection of basic speech units of Hindi using the above representation scheme, and (iv) Demonstration of a text-to-speech system for Hindi incorporating the coarticulation rules.

Chapter 1

INTRODUCTION

1.1 COARTICULATION AND ITS IMPORTANCE IN TEXT-T-O-SPEECH CONVERSION

Text-to-speech conversion involves synthesis of speech signal from the input text of a language. The nature of the synthesized speech is usually restricted to the normal reading style. While reading aloud a text, we use several knowledge sources which had been acquired by us over a period of time. These knowledge sources include low level knowledge pertaining to the features of human speech production mechanism and higher level knowledge such as intonation (fundamental frequency pattern), duration, stress etc. In order for the synthesized speech to be natural, these knowledge sources should be acquired and incorporated while synthesizing the speech from text. One such knowledge source is corresponding to the coarticulation phenomenon in continuous speech. Coarticulation refers to the influence of features of a speech unit on the features of another unit in continuous speech, which is caused by the physiological constraints of the vocal tract system. The research work described in this thesis is concerned with some issues in the design of a text-to-speech conversion system for Hindi, an Indian language, especially for incorporating the acoustic feature variations pertaining to coarticulation so as to make the synthesized speech sound intelligible and natural. The **coarticulation** is formulated as a knowledge base which consists of a set of rules. The representation and synthesis scheme of the speech units is designed to be flexible for modification by the coarticulation rules.

Knowledge is something which is not normally taught formally as a set of rules, but is acquired by human beings from examples, experience and practice. The knowledge, if formulated as a set of explicit rules, becomes incomplete, imprecise and inaccurate. One reason for this is our **inability** to express formally the constraints which govern the invisible system being studied. For example, a medical expertise formulated as a set of rules often fails because most of the governing constraints cannot be captured in the formulation. But if a system is constrained physically, it appears that the variability

of output due to knowledge of the constraints will be less. As an example, consider the writing process using cursive script which is analogous to the continuous speech. It involves pen, paper, hand and a set of symbols and hence the joining pattern of adjacent symbols is decided by the symbols and also by the physical medium. Coarticulation in continuous speech is similar to joining rules for the cursive script in several respects. Coarticulation decides the joining pattern between adjacent sounds in continuous speech. This causes the speech units to be different from what they are in isolation. Fig. 1.1 shows the sound spectrogram of the utterance /bho:jan/ (भोजन) and the constituent characters uttered in isolation. Coarticulation is a process involving simultaneous movement of several articulators and these movements are decided by the sounds being produced and the constraints due to physiological factors. Coarticulation patterns generally show consistency across speakers. This suggests that the coarticulation patterns can be formulated using examples of various contexts in natural speech. In our approach to text-to-speech system design, we use a collection of basic speech units and use the coarticulation knowledge formulated as a set of contextual rules to join these units to produce continuous speech output.

Besides acquisition of the coarticulation knowledge as a set of rules, another important issue is in incorporating this knowledge while synthesizing the speech. We need a representation of the speech units which has the flexibility of spectral and other parameter modification. Choosing appropriate speech parameters involves signal processing considerations. Construction of the basic unit data base involves speech data collection and analysis for the extraction of the desired parameters.

Following section describe the broad issues in building a text-to-speech system. These issues form the context of our discussion in the chapters to follow. The chapter concludes with a description of organization of the thesis.

1.2 ISSUES IN DEVELOPING A TEXT-TO-SPEECH SYSTEM

The issues in a text-to-speech system design are related to

- (i) Choice of the speech synthesis model
- (ii) Collection of data required for segmental synthesis
- (iii) Acquisition and incorporation of various knowledge sources required for producing natural sounding speech.

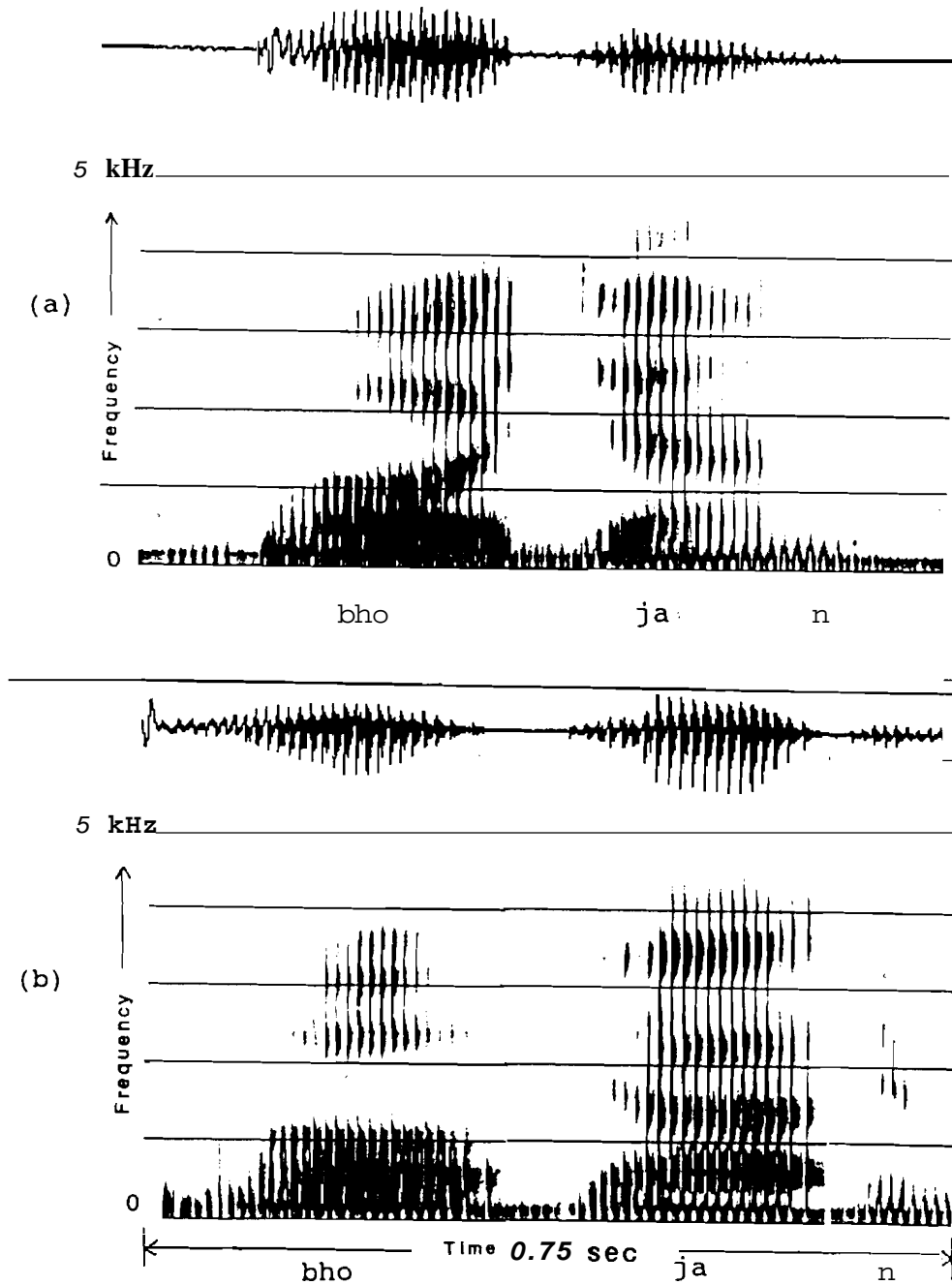


Fig. 1.1 Illustration of the **coarticulation** effect between speech segments

- (a) The waveform and spectrogram of the word /bhojan/ (भोजन) in Hindi
- (b) The waveform and spectrogram of the characters /bho/ (भो), /ja/ (ज) and /n/ (न) uttered in isolation. The changes in the spectrogram of the vowels of /bho/ and /ja/ in (a) above are due to the context.

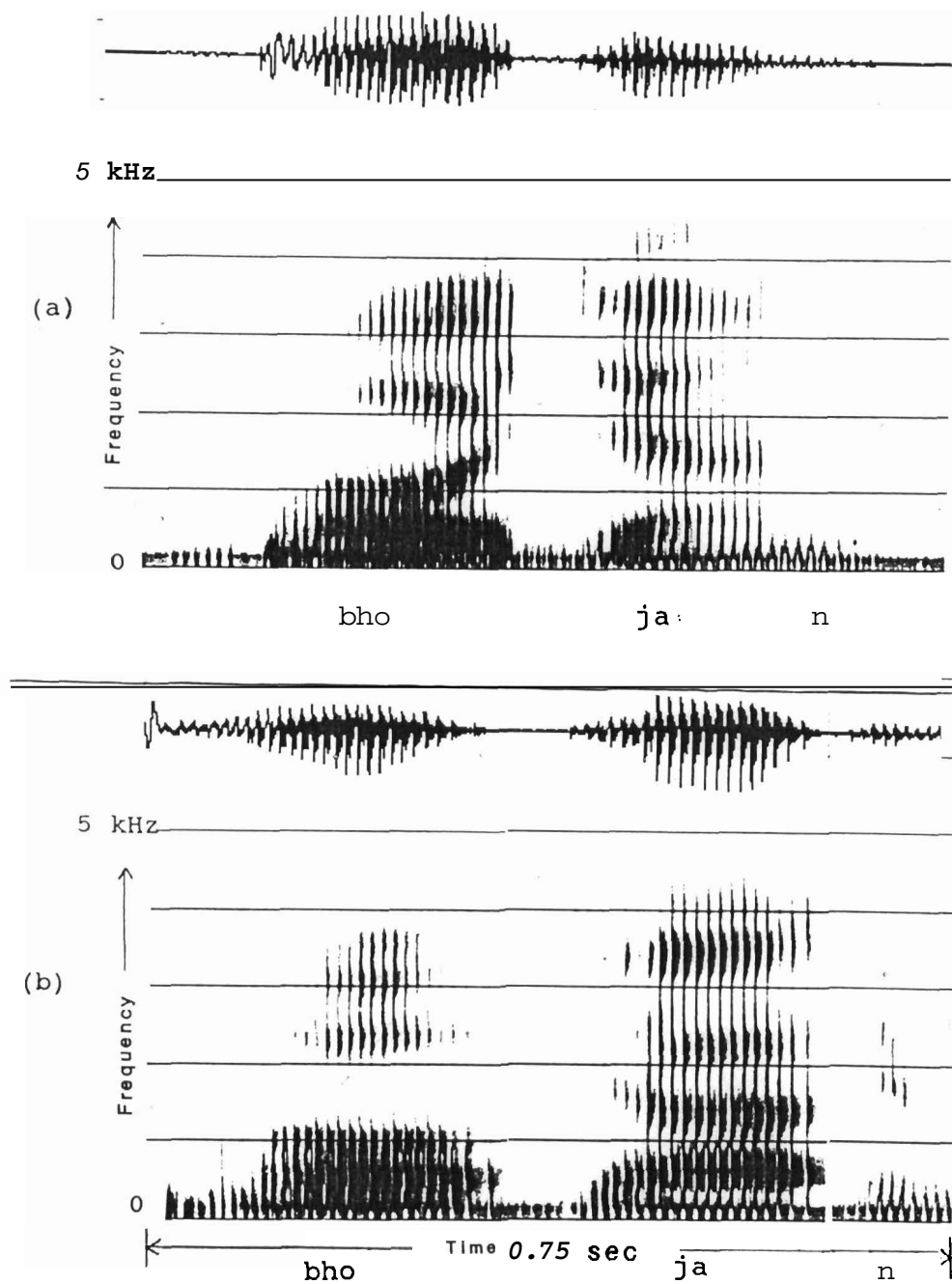


Fig. 1.1 Illustration of the coarticulation effect between speech segments

- (a) The waveform and spectrogram of the word /bhojan/ (भोजन) in Hindi
- (b) The waveform and spectrogram of the characters /bho/ (भो), /ja/ (ज) and /n/ (न) uttered in isolation. The changes in the spectrogram of the vowels of /bho/ and /ja/ in (a) above are due to the context.

1.2.1 Choice of the synthesis model

The common approaches used in a text-to-speech conversion system are: (i) Concatenation method and (ii) Synthesis-by-rule method. The concatenation method basically involves collecting and prestoring the basic speech units (basic units) (O'Shaughnessy, 1984; Lukaszewicz et al., 1987). In order to synthesize a given text, the text is parsed into a sequence of the constituent basic units, and then the corresponding segments are joined together to get the speech output. In synthesis-by-rule method a set of rules operate on the given text to produce a parameter sequence which is then synthesized (Allen et al., 1987; Allen, 1976; Docherty et al., 1988). The rules are for the generation of basic speech units (acoustic-phonetic knowledge) and for the naturalness of the synthesized speech (Holmes, 1988).

The concatenation method is simpler conceptually and requires less time for implementation. But it ^{is} not flexible for modification. The synthesis-by-rule method is flexible since no prestored speech is used for synthesis. But it requires both the knowledge for synthesis of basic speech units as well as the knowledge for naturalness to be acquired and incorporated. Also the inaccuracy in the knowledge appears as degradation in the quality of the synthesized speech.

1.2.2 Data collection for segmental synthesis

The second issue is concerned with the collection and representation of the information required for the synthesis of basic speech units. The way this is done depends on the synthesis model used. In concatenation method the speech data for all basic units of speech are collected from natural speech data and stored. But in synthesis-by-rule method the acoustic phonetic knowledge for synthesis of basic speech units is to be acquired and represented as rules or tables.

1.2.3 Knowledge sources for naturalness

Human speech is characterized by (i) segmental and (ii) suprasegmental features which collectively contribute to the naturalness of the speech. A segment refers to some small chosen unit of speech (eg. phonemes, syllables etc.). Segmental features refer to those which decide the phonetic quality of the segment. Suprasegmental or prosodic features have their domain extended over more than one segment i.e., syllables, morphemes, phrases, sentences etc. The suprasegmental

features are the rhythm (duration), stress (intensity) and intonation (pitch). The terms in parentheses are the acoustic parameters in which the corresponding features are manifested (Lehiste, 1970). Suprasegmental features are the overlaid functions of the corresponding segmental features (Rajesh Kumar, 1990). For example, intonation is the overlaid function of the periodicity (voicing) of all segments in an utterance. The suprasegmental features are influenced by factors such as phonetic and syntactic context, semantics, emotional state of the speaker etc.

In continuous speech, the segmental features are subjected to changes decided by the phonetic context. These changes are caused by the coarticulation effect and give rise to certain joining patterns between adjacent speech units.

The knowledge issue concerns with the acquisition of the knowledge pertaining to these features from natural speech and their incorporation into the text-to-speech system to make the synthesized speech sound intelligible and natural. The various categories of the knowledge, their domain, the acoustic parameters they affect and their linguistic function are given in Table 1.1.

Table 1.1 Categories of knowledge, their domain, parameters affected and linguistic function

Knowledge	Domain	Parameters	Linguistic function
Segmental	Speech unit	Spectral parameters	Phonetic quality of segments
Coarticulation	Speech units & transitions between them	Spectral parameters	Smooth flowing nature of speech
Suprasegmentals	Phrase, sentence etc.	Pitch, duration and intensity	Intonation, rhythm and stress

In this thesis we address some issues in the design of a text-to-speech system for Hindi using the concatenation model. We specifically address the issues in the acquisition and incorporation of the coarticulation knowledge.

1.3 ORGANIZATION OF THE THESIS

This thesis is organized as follows: Chapter 2 gives an overview of studies on coarticulation. It reviews studies on coarticulation knowledge for text-to-speech conversion for a few languages. Chapter 3 discusses the design details of a text-to-speech

system for Hindi. Chapter 4 discusses the issues in the acquisition and incorporation of the coarticulation knowledge into the text-to-speech system. Chapter 5 gives testing and evaluation procedures used to test the database of basic speech units and to evaluate the quality improvement in synthetic speech due to coarticulation rules. Chapter 6 gives a summary of this work.

Note on notation:

Throughout this thesis we use square brackets [] to enclose phonetic transcriptions and slashes / / to enclose phonemic transcriptions. The phonetic transcription of each of the consonants and vowels in Hindi is given in Appendix 5.

Chapter 2

STUDIES ON COARTICULATION - AN OVERVIEW

2.1 INTRODUCTION

In normal conversation we speak at a rate of 100 - 170 words per minute. Each word is, on an average, made up of around five different speech sounds. Faster rate of speech, say over 200 words per minute, is also possible. Yet these words are spoken with sufficient precision, though they could be different if spoken in isolation. Phonologically significant speech sounds have to be maintained distinctly in order to convey the correct meaning of the utterance.

High quality of speech output is possible while reading aloud a predetermined text consisting of different classes of speech sounds, by manipulating the complex speech production system. No single part is solely responsible for speech production. Speech is a simultaneous but systematically coordinated activity of several muscular systems that are attached to the different parts of the speech production system. The varying degrees of overlapping or simultaneous movement of different organs involved in the production of speech segment is called coarticulation.

Fig.2.1 shows parts of the vocal tract system involved in the speech production. We can illustrate as to how parts of speech production mechanism that are functionally independent of one another yet work in remarkably good coordination. The lips, tongue, velum, and larynx function almost independent of one another but in a closely coordinated sequence. Simultaneous articulations (double coarticulation) are possible by the speech production mechanism as for example, a closure by the bilabials and lingual constriction at different points in the vocal tract simultaneously as in /a:pka:/ (आप का). Similarly, voiced fricatives such as /z/ (ज़) are produced when vocal cords keep vibrating while turbulence is created at a constriction in the vocal tract.

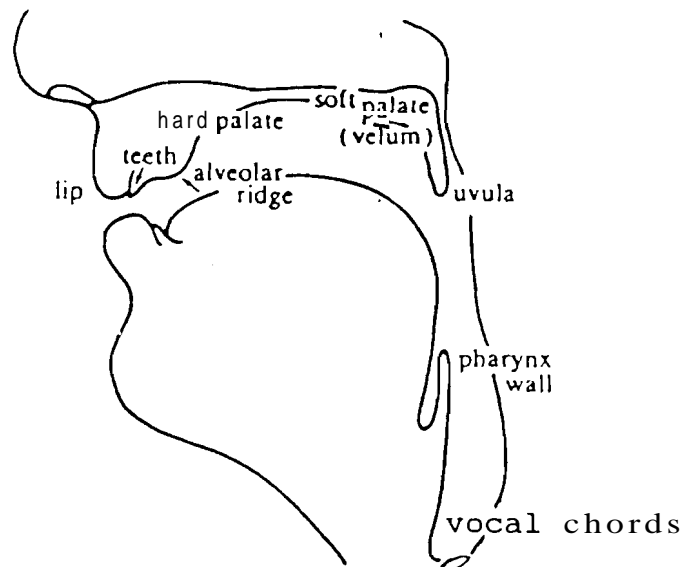


Fig. 2.1 Parts of vocal tract system involved in speech production

2.2 EVIDENCE FOR COARTICULATION

Articulatory, acoustic and perceptual evidences are readily available for explaining the coarticulation phenomenon. The articulatory evidence is obtained through the use of photography, or X-ray radiography. The muscle activity is studied through electromyography (EMG). Acoustic evidences are available through indirect means. The acoustic evidence has implications in the development of speech synthesis and speech recognition systems. Perceptual evidences indicate that the coarticulation phenomenon is useful for human recognition of the adjoining segments.

2.3 IMPORTANCE OF COARTICULATION

Speech production involves two fundamental aspects: stationary properties of phoneme realization and the dynamic properties governing the fusion of strings of discrete units (phonemes) into continuous speech. The transitional aspects of speech are as important as the stationary part for perception of speech. For example, isolated vowels are perceived in terms of locations of first two or three formants (vocal tract resonance frequencies), while the diphthongs are perceived due to vowel to vowel formant transition. Perception of vowels in typical context depends on the formant movements into, during and out of the vowel (O'Shaughnessy, 1987). It is reported that

vowel perception in CVC context is more aided by the CV and VC formant transitions than the steady region of the vowel itself (Strange et al, 1983). Studies using synthetic speech show that the acoustic cues to the perception of place of articulation feature of a consonant reside primarily in the spectral transition between the consonant and the adjacent vowel (Delattre et al, 1955). Formant transitions, in addition to helping phoneme identification, aid in auditory stream integration. Synthetic speech without such transitions tend to be perceived as two separate sound streams. Inadequate modeling of formant transitions may lead to fricatives being heard as isolated hisses superimposed on the rest of the synthetic speech (O'Shaughnessy, 1987).

2.4 TYPES OF COARTICULATION

There are two major types of coarticulation based on the direction of the coarticulation effects.

2.4.1 Anticipatory coarticulation or forward coarticulation

The gesture of the articulators for a given phoneme is set during the articulation of the preceding phoneme provided the gesture does not cause qualitatively significant changes in the current unit. Anticipatory coarticulation phenomena may even extend over many phonemes. This suggests that speech production is a planned activity. The coarticulatory effects of a given segment over a sequence of segments need not be same in the context of a different sequence of segments.

A striking example for the anticipatory coarticulation is the lip protrusion feature in French (Benguerel & Cowen, 1974). The feature lip protrusion may begin upto six segments in advance of its requisite appearance. This gesture of lip protrusion may be schematized as follows:

i s t r s t r y

•

The series of consonants beginning with /s/ after the unrounded vowel /i/ show the feature lip rounding in anticipation of the rounded /y/. In Hindi, anticipatory velar lowering for nasal consonant takes place during the preceding vowel. This can be shown as

p a: n

The vowel-to-consonant lingual ~-articulation in Hindi is also an example of anticipatory coarticulation.

2.4.2 Carry-over Coarticulation

The gesture of a given feature remain, to some extent, during the articulation of the following segment. Studies indicate that the carry-over coarticulation effects are due to the inertia of the articulators. Further, these coarticulation effects are attributed to low level phenomenon.

An example would be the presence of lip protrusion during the segment /s/ of the word 'boots' [buts]. The lip protrusion associated with /u/ has apparently been retained till the following segment /t/ and /s/ as in

b u t s
——→

Normally the carry-over effects do not extend beyond the immediately adjacent segment (Gay, 1977).

2.5 MODELS OF COARTICULATION

Several theoretical models have been proposed to explain the coarticulation phenomena. These models construe phoneme either as a bundle of features or articulatory targets which spread to the adjacent phonemes.

The overlapping of the articulatory movements definitely show that the speech organs are not capable of infinite acceleration. The transition between segments reveals the interactive influences of the segments. Coarticulation in cases where the influence exceeds the immediately adjacent segment cannot be explained by simple inertial effects of the muscles but calls for theoretical explanation. The motor controls during the intervening segments have simultaneous information about the segment ahead of as many as six segments. In these cases we can conclude that the coarticulation phenomenon exceed beyond two adjacent segments are well planned (Whalen, 1990).
in advance by the brain rather than caused by the inertia of articulators

Several models are available about the organizational unit based on the coarticulation knowledge.

2.5.1 Kozhevnikov-Chistovich articulatory syllable

Based on data in Russian, Kozhevnikov & Chistovich (1965) have argued that the range of forward coarticulation mirrors the articulatory unit. In their data the lip protrusion which is a characteristic feature of rounded vowel starts from the very first consonant in the sequence. This can be schematized as follows:

C (C)(C) V [rounded]

←—————

Thus, the articulatory syllable can be CV, CCV, CCCV etc. and it resembles Stetson's (1951) theory according to which CV is the basic articulatory unit. The articulatory command that begins at a consonant ends with the following vowel.

The disadvantage of the model is that the coarticulation between a vowel and the following segments are not considered. There are instances of anticipatory velar lowering during the articulation of a vowel in anticipation of a nasal consonant irrespective of a word boundary that exists in between (Moll & Daniloff, 1971). This anticipatory velar lowering cannot be explained by this model.

Ohman's (1966) studies based on spectrographic study of V₁CV₂ utterances reveals that anticipatory coarticulation ranges from V₁ to V₂. The acoustic characteristics of V₁ depends upon the characteristics of the transconsonantal vowel V₂. Thus the articulatory requirements of V₁ is determined by the articulatory requirements of V₂. According to this model the coarticulatory patterns would be

V₁ C V₂

Ohman's study is corroborated by a cinefluorographic study (Kent & Moll, 1972). This study reveals that the tongue position for [i] is slightly less in its height when followed by a stressed syllable with [u]. However, the tongue height for [i] is well maintained when it is followed by stressed syllable with a front vowel. This suggests that the articulatory characteristics of V₁ is conditioned by V₂.

MacNeilage & DeClerk (1969) compared the intensity of the coarticulation between CV and VC sequences on monosyllabic CVC data. This study suggests that the coarticulation between CV is more intricate than between VC sequence.

2.5.2 Wickelgren's context-sensitive model

In this theory speech units are coded allophonically as context sensitive elementary motor responses (Wickelgren, 1969). Each unit is encoded in the context of its left and right segments only. However, other models have shown that coarticulation phenomenon is not limited to only its immediately neighboring segments. Also the number of basic units of speech production goes beyond thousands as each segment is considered in the context of neighboring segments.

2.5.3 Feature-based models

This model assumes the input unit (phoneme or phonetic unit) is conceptualized as a vector of component feature, which are directly related to motor implementation. This model was shown in a computer simulation of speech articulation (Henke, 1966). It accepts a sequence of phonetic segments as input and gives the output as the instantaneous states of the articulatory system. Each articulator continuously moves to the next goal provided it does not affect the articulatory requirements of the adjacent units. Definitely, all the organs of speech are involved in the articulation of any intermediary segments look ahead for the segments in the string and move accordingly based on compatibility criterion. The compatibility criterion suggests that a feature is specified earlier than its present segment so long as the articulation by that feature does not contradict with the articulatory movements of the present segments concerned.

The features are binary valued, either “+” or “-”. When the feature is not relevant “0” is specified. At the articulatory goal level a specified feature of a segment sets in much ahead. For illustration, we can take the case of lip protrusion coarticulation in French (Bengueral & Cowen, 1974). For the segments in (1) below, the phonetic features of lip protrusion are specified in (2). The segment [i] is specified a “-” value for the feature, lip protrusion. The final segment /y/ is valued as “+”. Other segments in between are specified as “0”s as their values are not critical. Therefore, the anticipatory coarticulation of the feature lip protrusion takes place immediately after the preceding vowel which has a “-” lip protrusion feature as in (3).

(1)	i	s	t	r	s	t	r	y
(2)	-	0	0	0	0	0	0	+
(3)	-	+	+	+	+	+	+	+

An inadequacy of this model is that for more than 50% of the time the lip rounding feature starts during the articulation of the vowel [i]. Specification of a “+” value for the unrounded vowel prevents phonetic contrast in French.

2.5.4 Hierarchical model

This model assumes many levels of speech organization with complex interactions between them. In the model proposed by Liberman (1970) the features have specific articulatory muscle targets and the phones are organized into syllables. The syllables also have overlapping but independent articulatory components.

2.6 STUDIES ON COARTICULATION FOR TEXT-TO-SPEECH SYSTEMS FOR SOME MAJOR LANGUAGES

Studies on the development of text-to-speech systems for some major languages incorporating the coarticulation knowledge are discussed in the following sections.

A rule based segmental synthesis module for French, which makes use of coarticulation rules to improve quality, is reported (Rizet, 1991). The module converts an allophonic string into an acoustic parameter file to be sent to the synthesizer. A hybrid **cascade/parallel** synthesizer was chosen. The rule set is based on a phonetic feature approach. The feature matrix was **designed** with respect to the following phonetic description: Vowels are described using the **front/back**, **rounded/unrounded** and **oral/nasal** contrasts and four values on the **open/close** scale. Consonants are described using the place and manner of articulation and the **voiced/unvoiced** contrast. The default specifications of the control parameters for the phonetic segments are stored in a table. The parameters include formant values and their bandwidths, amplitude values (for parallel formant synthesizer) various durations like steady, left and right side transition durations of a unit. The table contains about 1200 target values extracted from a male speaker data base. The main module applies a coarticulation rule set using the default values in a table and produces the synthesizer control parameters

every 5 ms. A coarticulation rule specifies the changes to be made in the default parameters of a phonetic segment based on its left and right contexts which are one or more phonetic segments and/or phonetic features.

G. Heike et al, (1991) described modeling coarticulation for a German text-to-speech system KOLLE, which is based on the articulatory model of synthesis. He emphasizes the auditory feedback strategies, combined with general phonetic knowledge about the relation between articulation and acoustics as the basis for the generation of coarticulation rules. The KOLLE synthesis system produces a sequence of articulatory movements and a synthesized speech signal from a string of orthographic symbols input to the system. This is done by a pair of modules: the transformation module and the articulatory module. The input to the transformation module is a string of phonetic symbols derived from the original orthographic input string. The output of this module is a sequence of articulatory target positions, separated in time by 6.4 ms. This is input to the articulatory module which generates the area functions of vocal tract. The sequence of area functions is then used to synthesize the speech. The transformation module makes use of articulatory target positions associated with each input symbol. These articulatory target positions are initially defined as a set of articulatory parameters. In the case of consonants, the nondistinctive parameters are left free. These are to be filled depending on the neighboring symbols in the input string, physiological and dynamic restrictions of possible articulator movements and language specific restrictions. These restrictions form the basis of coarticulation rules. For the KOLLE system, auditory feedback strategies are used to generate rules. This auditory control focuses on segmental features such as quality of the intended segments (especially the consonants) and the correct number of segments (i.e, no additional or missing sounds).

A neural network based spectral interpolation for speech synthesis by rule has been reported (Ishikawa et al, 1991). Two types of artificial neural networks were used; one for phoneme recognition and the other for spectral synthesis. A recognition network performs mapping a spectrum on to a vector of which elements represent similarities to each phoneme (phonemic vector). A spectral synthesis network-performs inverse transformation of a recognition network, i.e., it maps the phonemic vector onto a spectrum. In order to interpolate two phonemic target spectra, we have to first obtain

the corresponding phonemic vectors. This is done by the recognition network. Input to this network is the normalized sampled mel-spectrum and the recognition result, which represents a phonemic vector, is obtained as an analog output. This neural network performs the inverse transformation of the synthesis network. Using these neural networks, interpolation between two synthesis units are performed. At first, applying the recognition network, two phonemic vectors are obtained from spectra at the last frame of the previous unit and at first frame in the following unit. Phonemic vectors of interpolation segment is calculated by linear or some other interpolation method between these vectors. Finally, the spectral synthesis network maps the phonemic vectors on to the spectra. Thus, the two types of networks perform the to and fro transformations between spectral and phonemic space. Since it is very difficult for these networks to carry out recognition or synthesis for all phonemes, consonants should be classified and multiple set of networks are required.

2.7 COARTICULATION KNOWLEDGE FOR A TEXT-TO-SPEECH SYSTEM FOR HINDI

The coarticulation phenomenon has to be captured and implemented for the identification of segments themselves besides giving naturalness to the speech output. The major coarticulation types in Hindi are discussed in the following sections.

2.7.1 Vowel-to-consonant coarticulation

Lingual coarticulation is concerned with the function of tongue as the active articulator. The movement of the tongue within the area of oral cavity modifies the cavity which is reflected in the resonance modes. The cavity changes are due to the movement of either the tongue or the lips towards the articulatory gesture of the ensuing consonant in a VC sequence.

2.7.2 Consonant-to-vowel coarticulation

Coarticulation takes place between the consonant and the following vowel and vice versa. The CV coarticulation is not our prime concern as ~~the coarticulation knowledge is preserved as~~ ^{it already represented in} the CV transition part ^{of} ~~is preserved in~~ the basic unit of the text-to-speech conversion system.

2.7.3 Vowel-to-vowel coarticulation

Vowels can be conveniently represented in the F₁-F₂ plane (Potter & Steinberg, 1950; Peterson & Barney, 1952; Fant, 1959). These F₁-F₂ correspond to the first two resonant modes of the vocal tract. These two formants are sufficient to synthesize satisfactorily the vowels. However, in continuous speech the dynamics of these formants are important for the perception of vowel in the given context. For appropriate representation of vowels in text-to-speech systems we need to know the information regarding the steady state values which are essential for the identification of the vowel. Studies show that the steady state values of vowels vary in continuous speech depending upon factors like the vowel space, the neighboring consonants etc.

In Hindi, vowel sequences of different types (quality) of vowels are common. A few examples are /hua:/ (हुआ), /a:ɑ/ (आओ), /ja:ie/ (जाइए) etc. The important information that is required for synthesizing the vowel transitions consists of the steady state formant values of vowels, the transition values and the duration of formant transition.

2.7.4 Nasalization

Nasalization is concerned with interaction of oral and nasal cavities in speech production. The velum guides the air that comes from the lungs either to the oral cavity by closing the nasal tract as in oral sounds, or the reverse as in nasal sounds, or an intermediate state as in the nasalized vowels. Variation in the acoustic data of nasals is caused by factors such as the width of the nasal cavity, the amount of mucous filled in the cavity, the hair in the nasal passage etc. (Fant, 1960).

Nasalization is a common coarticulation process in continuous speech. Vowels are nasalized in the context of nasal sounds and this process is rather a rule than exception. Nasal sounds that occur on either side influence the vowel. The degree of nasalization on the vowel depends upon the position of the nasal consonant. In most languages vowel is found to be more influenced when followed by a nasal consonant and therefore this feature (VC nasalization) is considered universal (Ferguson, 1975). However, studies are also available to show that the preceding consonant influences the vowel to a large extent (Nagamma reddy, 1990). In Hindi, vowels seem to be more influenced by the following consonant (Ohala, 1983). The amount of nasalization of the

vowel as in /na:m/ (नाम) is more when it is both preceded and followed by nasal consonants in Hindi. The information on quantity of nasalization is required to adequately modify the parameters concerned in speech synthesis systems.

Perceptual analysis indicate the vowels in Hindi are nasalized even when there is a nasal consonant across the intermediary speech segments [w, y, h, r] in a sequence as in [khaɖa:u:] (खड़ाऊँ) in which the velum starts lowering after the initial stop (Ohala, 1983).

Nasalization is studied from many perspectives - speech perception, speech synthesis and articulatory phonetics. (Smith, 1951; House & Stevens, 1956; Fant, 1960). Though there are differences in the observations of the above studies, there is agreement on some of the points on vowel nasalization. They are: (i) The intensity level of the F₁ is less than that of F₂ and (ii) The energy level of the formants is less than that of oral vowels. In our study on Hindi nasalized vowels, we have observed that the second formant frequency is slightly shifted towards right, particularly in the open and back vowels. This is shown in the Fig. 2.2. Also, the energy level of the second formant is slightly higher in the nasalized vowel.

In Hindi orthography there are two signs used for nasalization, *anunasika* (ँ) and *anmara* (ं). *Anunasika* is used for indicating nasalized vowels as in /hã:/ (हाँ). However, the sign *anmara* is also used to indicate nasalized vowels when there are strokes on the top of the letter as in /haĩ/ (हैं). Anusvara is normally used to indicate homorganic nasals word medially. The following are some of the examples.

/andar/ (अंदर)

/anda:/ (अंडा)

/si:ncna:/ (सीचना)

/a:nkh/ (आंख)

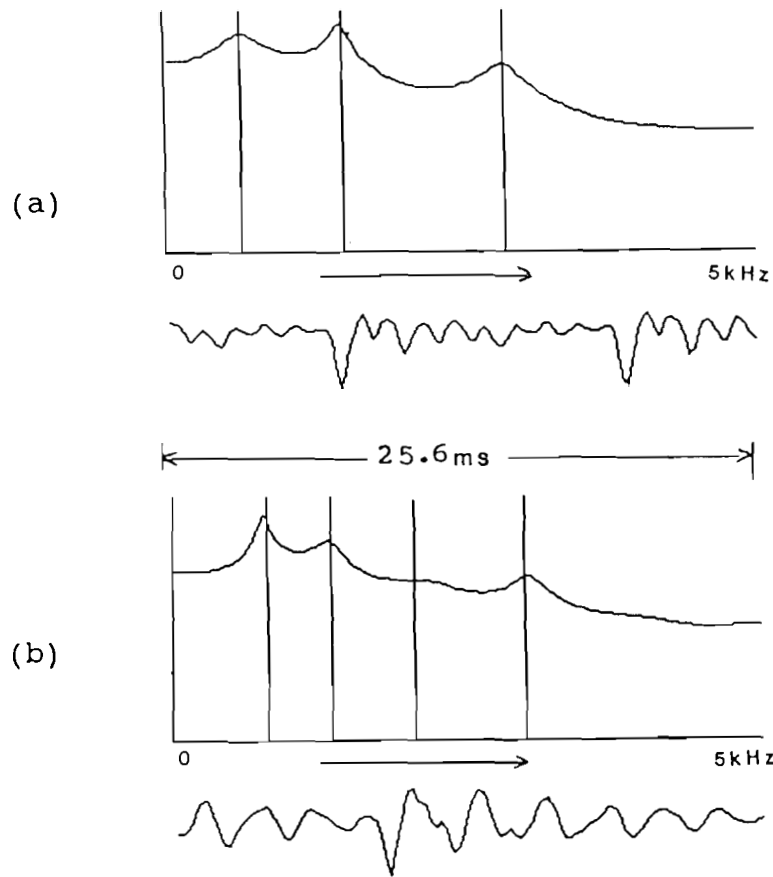


Fig. 2.2 (a) LP spectrum and waveform of nasalized vowel [ã]

(The vertical lines correspond to the formant peaks)

(b) LP spectrum for the oral vowel [a]

2.7.5 Consonant-to-consonant coarticulation

Coarticulation takes place between consonants (within clusters) as well. The following are our observations on the effect of coarticulation on consonant durations in cluster consonants: In stop consonant clusters of type C_1C_1 (i.e., both consonants are the same) the duration of the cluster is more than that of a single consonant. For example, in /bacca:/ (बक्का) the silence duration of /cc/ (क्क) is about 1.5 times that of a /c/ (क्) in word medial position. In consonant clusters of type stop followed by

glide (such as /khy/ (क्य) in /mukhya/ (मुख्य)) the duration of the stop consonant is increased about 1.5 times.

2.8 FACTORS AFFECTING COARTICULATION

The major factors that affect coarticulation are the language specific constraints and external factors like speaking rate and style. The language specific constraints include the proximity of phonologically contrastive units in a sequence (Manuel, 1990). In language with crowded vowel space, for example, the range of coarticulation is small. On the other hand, if the vowels are spread apart, the amount of coarticulation will be more. The prosodic and syntactic environments also affect the degree of coarticulation.

2.9 SUMMARY

Speech is a simultaneous but systematically coordinated activity of several muscular systems that are attached to different parts of the speech production system. The varying degrees of overlapping or simultaneous movement of different organs involved in the production of speech segment is called coarticulation. The transitional aspects of speech are as important as the stationary part for perception of speech. Several theoretical models have been proposed to explain the coarticulation phenomena. There are two major types of coarticulation based on the direction of the coarticulation effects, namely, anticipatory coarticulation or forward coarticulation and carry-over coarticulation. Coarticulation takes place between consonants (within clusters), between vowels and between the consonant and the following vowel and vice versa. The coarticulation phenomena is also classified in terms of the articulators responsible for it. Lingual coarticulation is concerned with the function of tongue as the active articulator. Nasalization is concerned with interaction of oral and nasal cavities in speech production. Vowels are nasalized in the context of nasal sounds and this process is rather a rule than exception. The major factors that affect coarticulation are the language specific constraints and external factors like speaking rate and style.

Chapter 3

ARCHITECTURE OF A TEXT-TO-SPEECH SYSTEM FOR HINDI

3.1 INTRODUCTION

The commonly used approaches to text-to-speech system design are: (i) Concatenation method and (ii) Synthesis-by-rule method. There are two schemes for concatenation method: the waveform concatenation and the parameter concatenation. We have chosen the parameter concatenation scheme since it gives flexibility for manipulation so that we can incorporate various knowledge sources to improve the naturalness of the synthesized speech. The parameter concatenation approach is described in section 3.2. Various phases of the text-to-speech conversion process are described in section 3.3. Section 3.4 examines various alternatives for the basic units for synthesis and gives the justification for choosing the characters of Hindi as basic units. It also discusses representation schemes for basic units with emphasis on the flexibility of parameter modification. Finally issues in constructing the data base of basic units in this representation are discussed.

3.2 SPEECH SYNTHESIS MODEL

In the concatenation approach to text-to-speech system design, the chosen basic units of speech can be stored using time domain coding or in the form of some parameters suitable for synthesis. The former method is called the waveform concatenation method and the latter is called parameter concatenation method. In the waveform concatenation scheme (Srikant et al, 1989) speech sounds of all basic units are sampled and prestored digitally. The basic units corresponding to the text to be synthesized are joined in the appropriate sequence to produce the speech. The speech output from this method does have high intelligibility but lacks naturalness and smooth flowing nature of speech. Limitations of waveform concatenation model are: (i) It is not flexible to facilitate the incorporation of various knowledge sources to improve the naturalness and (ii) It requires large amount of storage. Some of these

disadvantages are overcome in parameter concatenation model (Sriram et al, 1989) by using appropriate parameters. The parameter concatenation approach requires less storage but takes more computation time than the waveform concatenation method. Because of the flexibility the parameter concatenation model is chosen for the present system. The parameters used are the Linear Prediction (LP) coefficients or formant frequencies to model the vocal tract filter characteristics and the pitch and gain (intensity) to model the source (vocal chords excitation) of the human speech production system. The speech synthesis model using these parameters is shown schematically in Fig. 3.1.

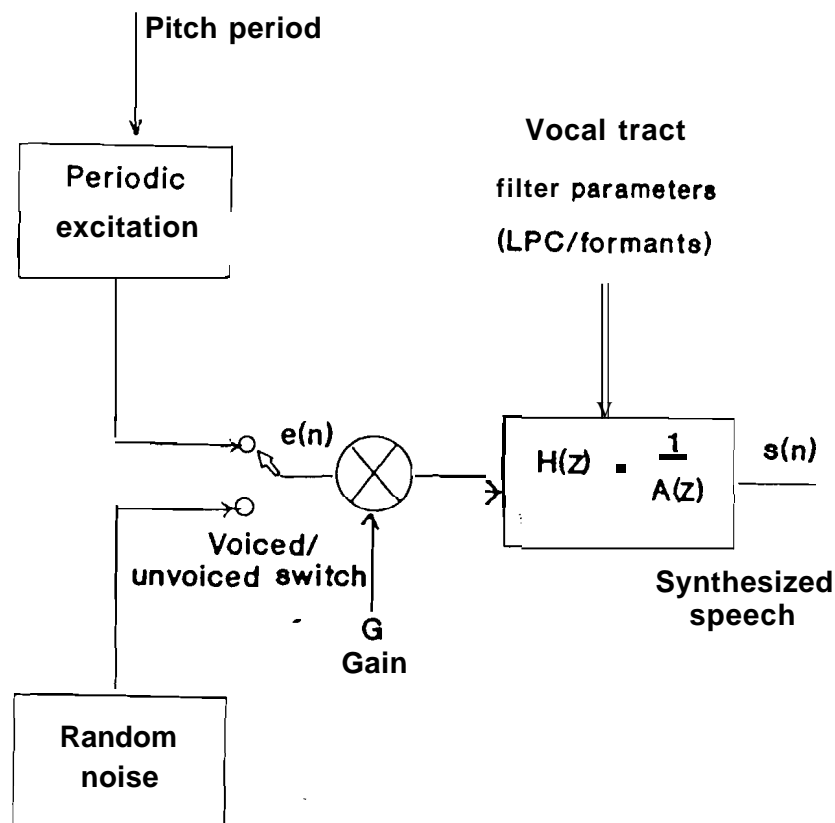


Fig. 3.1 Speech synthesis model

The vocal tract system is represented by the time varying digital filter whose spectral characteristics is specified by the LP coefficients or formant frequencies. The excitation signal is generated from the pitch and gain parameters using an excitation model. The basic units chosen are characters of Hindi for reasons explained in Section 3.4. The representation scheme of basic units uses LPCs to represent regions like voice bar (initial voicing in consonants), burst, aspiration etc. The vowel region which is mainly influenced by coarticulation is represented using formants. This is very convenient since coarticulation rules are formulated in terms of formant frequency changes. The excitation generator represents the source of the human speech production system. This is represented by periodic sequence of impulses in the case of voiced excitation and random noise for unvoiced excitation. The intensity is controlled by the gain parameter. This model is sufficient to generate all basic units except nasals for which we use an additional pole-zero pair in the filter.

3.3 VARIOUS PHASES IN TEXT-TO-SPEECH CONVERSION

The text-to-speech conversion process involves two major phases: (i) The text analysis phase and (ii) Synthesis phase. The former phase consists of preprocessing the input text to expand abbreviations etc. and then parsing it into a sequence of basic units of speech. The synthesis process involves the concatenation of the parameters of these units in the correct sequence and synthesis after the application of both segmental and suprasegmental rules for naturalness. The flow chart in Fig. 3.2 shows the various kinds of processing involved in the text-to-speech conversion.

3.3.1 Input

The input is Hindi text in Indian Standard Code for Information Interchange (ISCII). This may be typed in through an ISCII key board or input from a prestored ISCII file. The table of the ISCII code is given in Appendix 3.

3.3.2 Preprocessor

The function of the preprocessor is to expand the abbreviations in the text and also to convert the numerals to corresponding text. This is done usually with the help of look up tables. The preprocessor functions are listed below with examples.

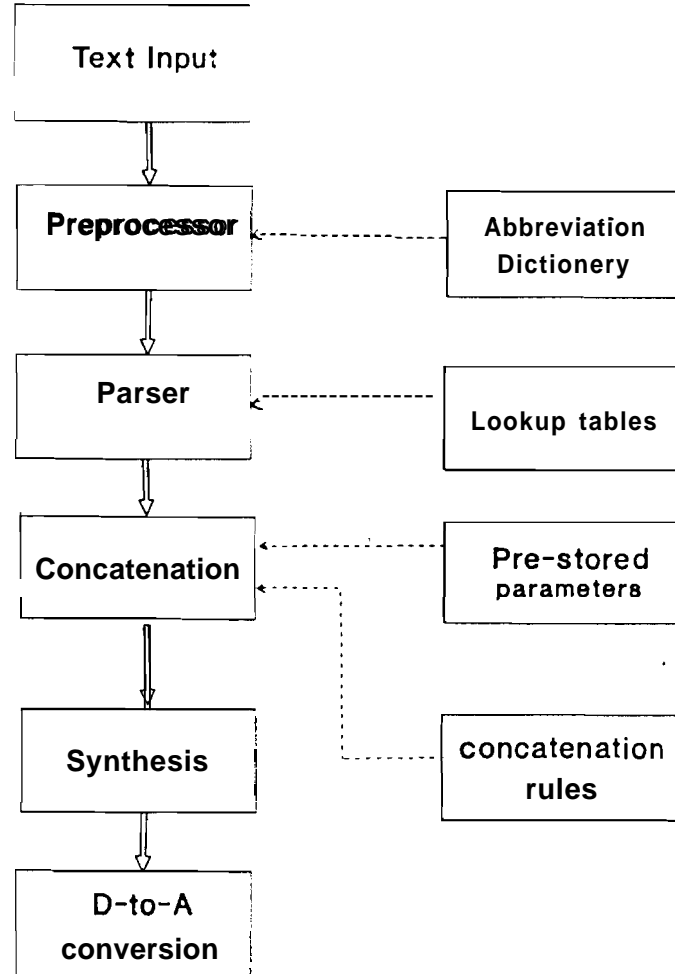


Fig. 3.2 Phases in text-to-speech conversion

- (i) Expand abbreviations to their spoken forms.

Eg. /da:n/ (डॉ) is expanded to /da:ktar/ (डाक्टर).

- (ii) Convert numerals to their text form.

Eg. 120.45 is expanded to /e:k sau bi:s dasamlav ca:r pa:nc/

(एक सौ बीस दशमलव चार पाँच)

- (iii) Convert year from numeral to text form.

Eg. /san 1947/ (सन् 1947) is expanded to /san unni:s sau sainta:li:s/

(सन् उन्नीस सौ सैंतालीस).

- (iv) Convert currency to their text form.

Eg. /ru: 12.351 (रु. 12.35) is expanded to /ru:paye: ba:rah paise: painthi:s/ (रुपये बारह पैसे पैंतीस).

The **preprocessor** makes use of two look up tables, one for abbreviations and the other for numerals (see Appendix 4).

3.3.3 Parser

The parser module parses the input text into a sequence of basic unit names. This module is simpler for Hindi as compared to nonphonetic languages like English (where letter-to-phoneme rules or dictionary lookups are used). In the present system name of a character (basic unit) consists of four components.

- (a) Consonant name (if present)
- (b) Vowel name (if present)
- (c) Nasalization indicator (if present)
- (d) Position indicator (isolated or within a word)

Each character in the input text has a corresponding sequence of ISCII codes. For all ISCII codes whose type is 'C' or 'V' their names are obtained from a lookup table (see Appendix 3). If any ISCII code is of type 'N', then the nasalized indicator is 'n'. Otherwise it is null. For all basic units occurring in polysyllabic words, the position indicator is '2'. If the basic unit occurs as a monosyllable word, the position indicator is '3'. The name of a basic unit is obtained by concatenating these components in the sequence listed above. In case of delimiters such as “,” or “|”, the name of the basic unit is a special string. The names used for the delimiters are given in Table 3.1.

Table.3.1 Basic unit names of the delimiters

Basic Unit	Name
,	comma
	bar
b	blank
!	exclam
?	qmark

There is an exception to the phonetic nature of Hindi. This is the vowel suppression at word final position and also in word medial position in some words. For

example, /kamala/ (कमल) is pronounced as /kama/ in Hindi and /karata:/ (करता) as /karta:/. The word final vowel suppression is built into the parser easily. The word medial case is not implemented because it is difficult to decide correctly from the context which vowel to suppress.

As an example, the sequence of ISCII codes: 90 108 99 82 32 107 97 108 99 108 32 84 115 104 32 107 116 32 121 (corresponding to the input text/bha:rat hama:ra: de:š haĩ/ or भारत हमारा देवा है ।) is parsed into the following sequence of basic units: *bhaa2, ra2, th2, blank, ha2, maa2, raa2, blank, dhe2, sh2, blank, hein3, bar.*

33.4 Synthesis of speech from parameters of the basic units

The input to this module is a sequence of basic units. The parameters of the basic units are concatenated (The representation of basic units will be described in Section 3.4). The resulting pitch contour is manipulated to incorporate prosodic information, namely intonation (Madhukumar et al, 1992). The durations of basic units are modified depending on the context in which they occur (Rajesh Kumar, 1990). Appropriate transition patterns are given between adjacent basic units to bring in the coarticulation effects. The excitation signal is generated from pitch and gain contours. The nature of the excitation signal also decides the quality of the synthesized speech. One of the following models can be used for generating the excitation signal:

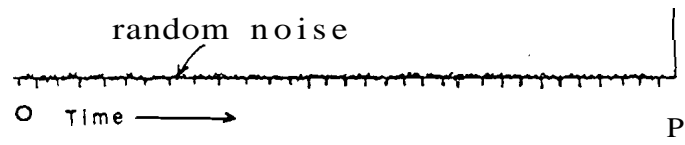
- (i) Impulse excitation
- (ii) Double impulse excitation
- (iii) Fant's excitation

Fig. 3.3 shows the excitation signal for a pitch cycle for these excitation models.

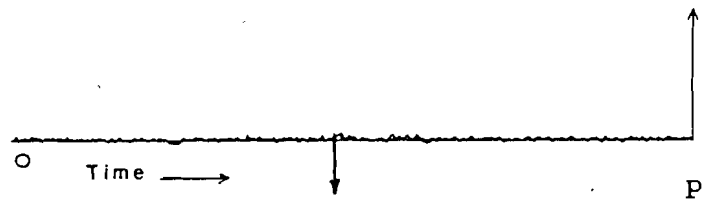
(i) Impulse excitation

A sequence of impulses, spaced at pitch intervals form the impulse excitation (Sriram et al, 1989). The amplitude of any impulse is decided by the gain at that instant. In order to remove the bias due to the positive impulses, some small negative impulses are introduced randomly in such a way that the average value is zero over any pitch period. In impulse excitation, The energy is concentrated at the locations of the impulses only and because of this the synthesized speech is slightly metallic.

(a)



(b)



(c)

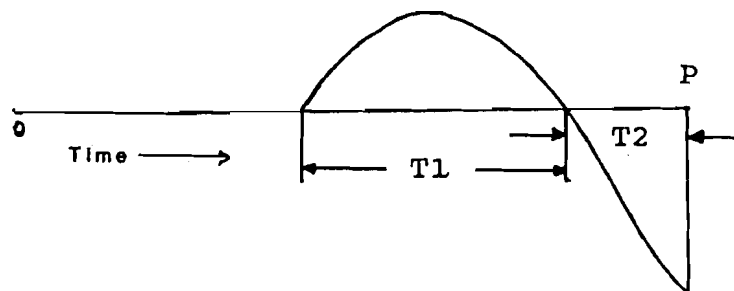


Fig. 3.3 (a) Impulse excitation
(b) Double impulse excitation
(c) Fant's excitation model

(ii) Double impulse excitation

It is similar to the impulse excitation except that the excitation is a sequence of pairs of impulses of opposite sign which are located at certain percentage positions of the pitch interval (Sriram et al, 1989). In this case, the energy in each pitch cycle is distributed over the two impulses.

(iii) Fant's model for excitation

The Fant's excitation models the glottal excitation in human speech (Yegnanarayana et al, 1984; Childers et al, 1989). In this model, The energy is distributed over the pitch period. Hence the quality of the resultant speech is significantly better when compared to speech synthesized using any of the impulse excitations.

3.4 BASIC UNITS AND THEIR REPRESENTATION

The core of a text-to-speech system based on concatenation model is its inventory of basic units of speech. The first step in designing such a text-to-speech system is to decide as to what should be the basic units. The choice of basic unit involves tradeoffs such as space-time, intelligibility-flexibility etc. (Rajesh Kumar, 1990; Allen et al, 1987). The second design issue is to find an appropriate representation for the units. This also involves conflicting requirements of space-time, quality-flexibility etc. Section 3.4.1 discusses various tradeoffs and options involved in choosing the basic unit. The representation scheme of basic units is given in detail in section 3.4.2. Section 3.4.3 examines the issues in the construction of the basic unit database in the representation scheme.

3.4.1 Choice of Basic unit for Synthesis

Choosing the proper basic unit type is an important issue because (i) size of the database (ii) quality of synthesized speech and (iii) computational time depend on the basic unit type. All these factors are interrelated and therefore choosing the basic unit involves tradeoffs among them. One such tradeoff is between the size of the vocabulary and the quality of synthesized speech. As the size of the basic unit goes up from phoneme to sentence the size of vocabulary increases nonlinearly making it practically impossible to collect all basic units. But the intelligibility and naturalness of

speech increases with increasing size of the basic unit. Also the computation in synthesis is less for larger basic units. Another **tradeoff** is between the size of the basic unit and the knowledge required for synthesis. **The** synthesis-by-rule approach has no basic units but the acoustic-phonetic knowledge required to generate them is needed. Synthesis using phonemes requires knowledge of **coarticulation** and prosody while using prestored words or phrases requires mainly the prosodic knowledge.

The basic unit of the vocabulary can be any one of these: (i) Sentence (ii) Word (iii) Syllable (iv) Character (v) **Diphone** and (vi) Phoneme. Each has got its own advantages and disadvantages. In the context of an unrestricted text-to-speech system, sentence cannot be a good choice for the basic unit since it is not practical to enumerate and **prestore** all sentences in a language. Word also is not a proper choice since the number of words to be stored is large. Syllable does not have a precise definition in terms of its constituents to arrive at the inventory of syllables of a given language. Diphones, though they capture transitions between phonemes, are large in number (all CV and VC combinations) and also they cause discontinuity in steady regions like vowels. Phonemes, though they are only about forty in number, require correct knowledge and procedures for concatenation to produce intelligible units. The knowledge is difficult to acquire and the concatenation procedures are difficult to implement in practice.

The basic unit type we have chosen is a character. A character in an Indian language can be any one of (i) consonant vowel sequence (CV) (ii) consonant alone (C) and (iii) vowel alone (V). Cluster characters CCV, CCCV etc. can be generated by combining appropriate Cs and a CV. The advantages in choosing the character are: (i) Indian languages are phonetic in nature. This means each character has got a unique script and pronunciation. This is unlike English and similar languages for which this correspondence does not exist, necessitating the use of a pronunciation dictionary (ii) The number of characters also is not too large (about 350) (iii) Consonant to vowel transitions are preserved in characters (iv) Characters have got more natural pronunciation than the phonemes, **diphones** etc.

3.4.2 Representation of basic units

It is desirable to have a representation scheme for basic units which requires less storage space and less time for synthesis and is flexible enough for incorporating various segmental and suprasegmental effects for naturalness. It is difficult to **satisfy** all these requirements since they are conflicting in nature. The space-time **tradeoff** is that if we want to reduce the space requirements by using some coding of the basic units, the time required for synthesis is more. The waveform concatenation model which requires maximum space is the fastest (real time) of all synthesis methods. Yet another **tradeoff** is between flexibility of representation and intelligibility of synthesized speech. In waveform concatenation, we have maximum intelligibility but no flexibility whereas in the synthesis-by-rules system, there is maximum flexibility, but due to the absence of the required knowledge for **synthesis**, the intelligibility is not satisfactory. The representation used for basic units makes use of several parameters derived from natural speech data. It is not just a simple coding of the speech data of basic units using coding methods like LPC, since they are not flexible. The following paragraphs discuss the representation scheme.

In section 3.2 we have outlined in brief the synthesis model used in the present system. The representation is based on a model of the human speech production mechanism (**Papamichalis, 1987**) (see Fig. 3.1). In this a time varying filter which models the vocal tract is excited by an excitation signal which approximates the air flow characteristics of the glottis, the excitation generator of the human voice production system. The vocal tract system is modeled by an all-pole filter. The speech **coding/synthesis** methods based on this model are: (i) The Linear Prediction (LP) synthesis and (ii) The Formant synthesis.

3.4.2.1 Linear Prediction synthesis

The Linear Prediction method (Makhoul, 1975) is one of the most popular approaches for coding speech. It is based on the speech production model shown earlier. The vocal tract is modeled as an all-pole digital filter, i.e., as a filter that has only poles and no zeros. Incorporating the gain, G , into the filter, we can express it as

$$H(z) = \frac{G}{1 + a_1 z^{-1} + \dots + a_p z^{-p}} = \frac{S(z)}{E(z)}$$

where p is the order of the model. If $s(n)$ is the speech output of the model and $e(n)$ is the excitation input, the equation above can be written in the time domain as

$$S(n) = G e(n) - a_1 s(n-1) - \dots - a_p s(n-p)$$

In other words, every speech sample is computed as a linear combination of the previous speech samples with a contribution from the excitation. The coefficients a_k , $k=1, p$ are to be computed for each frame of speech data using the autocorrelation normal equations (Rabiner et al, 1978), which are derived by minimizing the total error energy. The gain G is computed for each frame from the minimum error energy. The advantages of the LPC model are the following:

- (i) It can represent fairly most of the speech sounds except nasals and voiced fricatives.
- (ii) Efficient methods for automatic extraction of LP coefficients are available.

The disadvantages are the following:

- (i) It cannot model zeros in the speech spectrum and hence it does not model speech sounds like nasals satisfactorily.
- (ii) It is very rigid against spectral level modification.

Hence adjustments of peak frequencies (formants) in the spectrum are impossible. The second disadvantage is very severe making it impossible to incorporate the knowledge due to coarticulation which modify the spectral peaks contextually. So we have to think of other ways of representing the vocal tract filter for synthesis. The formant synthesis is one such alternative.

3.4.2.2 Formant synthesis

It is known that, perceptually the most important elements of speech spectrum are the energy concentrations (peaks) around frequencies which correspond to the vocal tract resonances, also called formants. The strength or energy content of a formant is decided by its bandwidth. Typically, the first three formant frequencies are sufficient to represent the important information in the spectrum of vowels. There are two models of speech synthesis based on formants: (i) the cascade configuration and (ii) the parallel configuration (Klatt, 1980; Holmes, 1983). In both cases a spectral

peak at formant frequency F and of bandwidth B is modeled using a second order all-pole filter given by

$$H(z) = \frac{G}{1 - a_1 z^{-1} - a_2 z^{-2}}$$

$$\text{where } a_1 = 2e^{-\pi BT} \cos(2\pi FT)$$

$$a_2 = -e^{-2\pi BT}$$

$$G = 1 - a_1 - a_2$$

The constant T is the sampling interval. The frequency response is obtained by evaluating the function along the unit circle in the z -plane (i.e., by setting $z = e^{j2\pi fT}$ in the expression for $H(z)$). In the cascade configuration, a chain is formed by connecting the output of a filter to another. The first one in the chain is then excited and the output is taken from the last. The transfer function of a cascade of n formants F_1, \dots, F_n with bandwidths B_1, \dots, B_n respectively is the product of transfer functions of individual formants

$$H(z) = \prod_{i=1}^n \frac{G_i}{1 - a_{i1} z^{-1} - a_{i2} z^{-2}}$$

The advantage of the cascade connection is that the relative amplitudes of the formant peaks for vowels come out just right without the need for individual amplitude controls for each formant (Allen et al, 1987). The disadvantage is that one still needs a parallel formant configuration for the generation of fricatives and plosive bursts - the vocal tract transfer function cannot be modeled adequately for these sounds by a cascade of resonators.

In the parallel configuration, all second order filters are excited by the same excitation signal and their outputs are summed to get the net output. Hence the transfer function is the sum of individual transfer functions. The peculiarity of the parallel configuration is the additional flexibility in controlling amplitudes of formants independently which makes them superior to the cascade configuration in fricative and plosive-like sound synthesis. But this is an overhead for vowel synthesis and is not required for cascade configuration.

We use the **LP** and the cascade formant synthesis models for the basic unit representation and synthesis. The excitation component of the basic unit is represented by the parameters pitch and gain. Pitch gives the period of voicing for voiced speech frames and is zero for unvoiced frames. The gain parameter specifies the scale factor using which the excitation signal is scaled before used for synthesis. The following considerations apply to choosing the proper combination of the above parameters to code a basic unit. The consonant region in general is not steady and does not have any formant structure. So we have chosen to use **LP** coefficients to represent them. But some consonants like semivowels are synthesized using formants. Vowels and vowel regions of CV units are represented using the first few formants (3 or 4 in number) and their bandwidths. The transition from C to V in CV units is prestored as changes in formant values and the transition time. In some cases where the transition is not clear or not visible as changes in formant frequencies (transition regions may coincide with aspiration or frication and hence may not be clearly visible as formant transitions), we include the transition region in the LPC represented portion. In all these cases, the CV transitions are captured in the basic units themselves. Since the vowels are represented using formants, we have the flexibility to modify the formants towards the end of the vowel. This is very important for the incorporation of the coarticulation knowledge.

The basic unit is encoded using information of two types: (i) Various speech parameter values and (ii) Control information - control frame numbers, durations etc. The basic unit description for each unit is stored in a table called *cvtable*. Components of the basic unit representation are:

- (i) Formant values and their bandwidths
 - (a) F_{i1} , **F_{i2}** , F_{i3} , F_{i4}
 - (b) F_1 , F_2 , F_3 , F_4
 - (c) F_{d1} , **F_{d2}** , F_{d3} , F_{d4}
 - (d) B_1 , B_2 , B_3 , B_4
- (ii) Control frame numbers
 - (a) Start_frame_no, No_of_frames
 - (b) Burst_frame_no, Onset_frame_no
 - (c) Pitch_frame_no1, Pitch_frame_no2
- (iii) Gain and pitch values
 - (a) Burst_gain, Consonant_gain, Vowel_gain
 - (b) Pitch_initial_voicing, Pitch_vowel
- (iv) Control durations
 - (a) CV_transition_duration
 - (b) Diphthong_steady_duration,
Diphthong_transition_duration

Fig. 3.3 shows some the control fields with respect to a basic unit waveform. The meaning of various fields are explained in the following.

(i) Formant values and their band widths

- (a) F_{i1} , **F_{i2}** , F_{i3} , F_{i4}

Values of formants at the onset of the vowel in a basic unit.

- (b) F_1 , F_2 , F_3 , F_4

Values of formants in the steady state region of the vowel in a basic unit.

- (c) F_{d1} , **F_{d2}** , F_{d3} , F_{d4}

These fields are used by basic units whose vowel region is a diphthong. They represent the terminating formant values of the diphthong vowel. Diphthong units

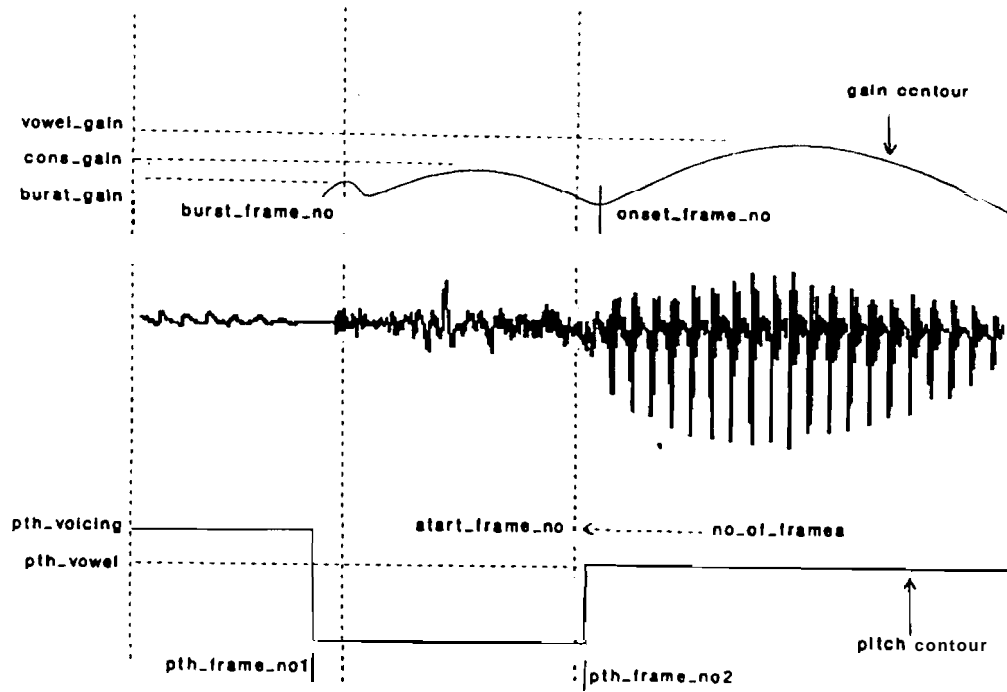


Fig. 3.4 Basic unit representation information with respect to waveform

are distinguished (by the system) from other units using the field dipthong-transition-duration. This field is nonzero for diphthongs only.

(d) **B₁, B₂, B₃, B₄**

Bandwidths of formants.

(ii) **Control frame numbers**

(a) **Start_frame_no, No-of-frames**

First one is the frame number at which switching from LP to formant synthesis takes place. This means all frames upto, but not including, start-frame_no will be represented using LPCs. The second field gives the number of frames represented using formants. The total number of frames in the basic unit is the sum start-frame_no - 1 + no-of-frames.

(b) **Burst-frame_no, Onset-frame-no**

The first field gives the frame number which contains the burst, if there is one. if the basic unit does not have a burst, this field is zero. **Upto** this frame number it is silence for voiceless stops and initial voicing for voiced stops. The second field is the onset frame number of the vowel.

(c) Pitch_frame_no1, Pitch_frame_no2

These fields mark respectively the frames at which initial voicing (of the consonantal region) ends and final voicing (of vowel region) begins in a CV unit. For an all voiced unit, these will be adjacent frames. The region between these two frames is unvoiced.

(iii) Gain and pitch values

(a) Burst_gain, Consonant_gain, Vowel_gain

The first field specifies the gain value to be used for the synthesis of the burst region. The second field gives the gain value used for the synthesis of the region between the burst and vowel onset (It may be the regions of aspiration or frication if present or the initial voiced regions of semivowels and nasals). The third field specifies the energy (gain) for vowel region. All these values are expressed as percentages of a fixed maximum value. These values, along with the two control frame numbers burst_frame_no and onset_frame_no, are used to fit a gain contour required for the synthesis of the basic unit. This saves space, time and a lot of hand editing which would be needed otherwise if we stored the gain contour as such for all basic units.

(b) Pitch_initial_voicing, Pitch_vowel

The first field is the pitch value of initial voiced region and the second is that of the vowel region of the basic unit. These, along with the two pitch frame numbers described above, decides the pitch contour of the basic unit.

(iv) Control durations

(a) CV_transition_duration

It is the duration in frames, of the formant transition in the beginning of the vowel region. This along with the initial and steady formant values, decides the CV formant transition.

(b) Diphthong_steady-duration, Diphthong_transition-duration

The first is the duration in frames of the steady region, if present, in the vowel after the CV transition. The second one is the duration of transition to the final formant values of the diphthong vowel. The fields are zero for nondiphthong units.

3.43 Construction of the basic unit data base

The issues in obtaining basic units in the above representation format are: (i) Collection speech data, (ii) Analysis of the speech data to extract the parameters required for filling the various fields of the basic unit representation and (iii) Coding the basic unit.

3.43.1 Data collection

The speech data is collected for each of the basic units according to certain guidelines. The data is collected for words containing the basic units (carrier words). The carrier words were spoken in isolation by a native Hindi speaker. Regarding the choice of carrier words there are two options: (i) meaningful words and (ii) nonsense words, We have selected nonsense words because of the following reasons (Rajesh Kumar, 1990):

- (i) The nonsense words, being unfamiliar to the speaker, are less subjected to undesirable prosodic bias introduced subconsciously by the speaker.
- (ii) The nonsense words allow us to quickly form a suitable carrier word to make the extraction of basic units easier.

In order to incorporate the prosodic features during synthesis, it is desirable that the stored basic units be devoid of any prosodic bias. This implies that the carrier words must be uttered with a flat intonation. It is also desirable that the units be devoid of any coarticulation effects since these are to be incorporated by contextual rules. The choice of carrier words should be such that the durations of basic units are affected less so that they can be used as the default (inherent) durations while activating the durational knowledge.

Regarding the format of carrier words used, some guidelines are used based on the above observations. Each carrier word contains 3 characters in the form C₁ C₂ C₃. For CV units, C₁ is any character, C₂ is the desired basic unit and C₃ is an unvoiced

stop consonant. For example, /kama:t/ (कमट) is a carrier word for /ma:/ (म). For C (consonant alone) units, C₁ is any character, C₂ is a CV type character in which C is the desired basic unit and V is the vowel /a/ (अ), and C₃ is an unvoiced stop. For example, /kamat/ (कमट) is a carrier word for /m/ (म). For V (vowel alone) units, C₁ is the desired basic unit, C₂ is a CV character where C is an unvoiced stop and C₃ is any character. As an example, /auka:t/ (औकट) is a carrier word for /au/ (औ). Some exceptions to the above are followed in the following cases:

- (i) Basic units in which C is /r/ (र):

It is difficult to extract /r/ (र) in the word medial position since its initial voiced region merges with the preceding vowel. So these are extracted from carrier words with the unit in the initial position. In this case C₁ is the desired basic unit, C₂ is an unvoiced stop character and C₃ is any character. For example, /ra:kat/ (राकट) is a carrier word for /ra:/ (रा).

- (ii) Basic units in which C is /h/ (ह):

The reason for choosing /h/ in word beginning is that its unvoiced region becomes too short in word medial position. The carrier words used have C₁ as the desired unit, C₂ an unvoiced stop character and any character as C₃. For example, /hikat/ (हिकट) is a carrier word for /hi/ (हि).

3.43.2 Analysis of the speech data

The carrier words recorded are digitized at 10 kHz sampling rate and are then stored as waveforms. Waveforms of the basic units are carefully separated from the waveforms of carrier words. The above two steps were done using an interactive speech digitizer/editor package. These waveforms are then subjected to the following analysis:

- (a) Extraction of LP Coefficients

The LP Coefficients are computed using the autocorrelation method, using the following values: LP order = 14, frame size = 25.6 msec. and shift between adjacent frames = 6.4 msec.

- (b) Extraction of formant frequencies

A method of formant extraction based on the group delay function was used (Yegnanarayana et al, **1990**; Yegnanarayana et al., **1991**). The first three formants are used in the basic unit coding. The frame size and shift values are same as those used for the LP analysis.

(c) Extraction of pitch

A method of pitch extraction based on the group delay function was used (Yegnanarayana et al, **1992**).

Since we use a model for the gain contour in the basic unit representation, the gain extracted from the basic unit waveform is not used. The automatic extraction of reliable bandwidths of formants is not feasible from the speech signal. The bandwidth values for the formants are chosen based on heuristics.

3.433 Coding of Basic Units

The basic units are coded using the parameters extracted and also setting the control frame numbers and durations appropriately for each unit. Since the representation is not uniform, this was done manually for each basic unit. A utility program (see Appendix 2) was developed to create and edit the basic unit representation. It allows to modify the contents of representation and test it by synthesizing the unit. Table 3.2 lists the contents of various fields of the representation for the basic unit /ka/ (ക).

Table 3.2 Basic unit representation information for /ka/ (क)

Name of the unit: ka2

$$1) F1 = 605 \text{ Hz} \quad 5) Fi1 = 546 \text{ Hz} \quad 9) Fd1 = 0 \text{ Hz}$$

$$2) F2 = 1200 \quad 6) Fi2 = 1445 \quad 10) Fd2 = 0$$

$$3) F3 = 2285 \quad 7) Fi3 = 2363 \quad 11) Fd3 = 0$$

$$4) F4 = 0 \quad 8) Fi4 = 0 \quad 12) Fd4 = 0$$

$$13) CV_trans_dur = 6 \quad 16) Start_frame-no = 21$$

$$14) Diphth_s_dur = 0 \quad 17) No_of_frames = 15$$

$$15) Diphth_t_dur = 0 \quad 18) Burst_frame-no = 18$$

$$19) Onset_frame-no = 21$$

$$20) Cons_gain = 1 \quad 23) Pitch_frame-no1 = 0$$

$$21) Burst_gain = 5 \quad 24) Pitch_frame-no2 = 21$$

$$22) Vowel_gain = 100 \quad 25) Pitch_init_voic = 0$$

$$26) Pitch_vowel = 71$$

$$27) Bandwidth1 (B1) = 150 \text{ Hz}$$

$$28) Bandwidth2 (B2) = 200$$

$$29) Bandwidth3 (B3) = 200$$

$$30) Bandwidth4 (B4) = 200$$

The basic unit is divided into two broad regions by the field start-frame-no. Frames 1 through start-frame-no - 1 are coded using the LPC which is stored in a file. These frames form the consonantal region of the basic unit which, for the basic unit /ka/, consists of the initial silence followed by burst. Frames start-frame-no to the

last frame are represented using formants and their bandwidths which are specified by various **fields** to be explained. These frames form the vowel **part of** the unit. The **no_of_frames** field gives the number of frames represented using formants. Thus total number of frames of the unit is $\text{start_frame_no} - 1 + \text{no_of_frames}$. The formant values **F₁,...** and **F₁,...** together with the **CV_trans_dur** collectively represent the vowel region. **F₁,...** are the onset formant values and the **F₁,...** are the steady formant values of the vowel. The duration of the CV transition is given by the **cv_trans_dur** as the number of frames in the transition region. The transition starts at frame **start_frame_no** with formant values **F₁,...** and ends at frame $\text{start_frame_no} + \text{cv_trans_dur} - 1$ with formant values **F₁,...**. The bandwidths of the formants are given by **B₁,...**. The gain contour of the basic unit is divided into various regions. The field **burst_frame_no** gives the frame number containing the burst. The gain of the burst is specified by **burst_gain**. The frames 1 through **burst_frame_no** represent the initial voicing or silence region. For the basic unit /ka/, this region is silence, so the system will assume a gain of zero (for voiced units, a small nonzero gain value will be used). The **onset_frame_no** gives the vowel onset frame number in the gain contour. The **cons_gain** field gives the gain of the consonant region between the burst and onset. For /ka/, it is very small since this region is of low energy. The **vowel_gain** specifies the maximum gain of the vowel. The two control frame numbers **burst_frame_no** and **onset_frame_no** together with the gain values **burst_gain**, **cons_gain** and **vowel_gain** specify the first level gain contour of the basic unit which would be modified later by contextual rules. The frame numbers specified by the fields **pitch_frame_no1** and **pitch_frame_no2** mark the end of the initial voicing (for voiced consonants) and the beginning of voicing of the vowel region, respectively. **pitch_initvoicing** gives the pitch period of initial voicing and **pitch_vowel** gives the pitch of vowel region. For /ka/, **pitch_initvoicing** is zero, being unvoiced. The frames between **pitch_initvoicing** and **pitch_vowel** are treated unvoiced by the system. Fields **Fd₁,...** are additional formant frequency values used for units containing diphthong vowels. **Diphth_steady_dur** and **diphth_trans_dur** specify, respectively, the initial steady duration and the transition duration of the diphthong vowel. Since /ka/ does not contain a diphthong vowel, these fields are set to zero.

3.5 SUMMARY

The first design issue in developing a text-to-speech system for an Indian language is the choice of synthesis model. The concatenation model was chosen for our text-to-speech system for Hindi. The next decision to be taken is about the choice of basic units and their representation. The characters of Hindi were chosen as the basic units. Various phases of the text-to-speech conversion process are the preprocessing, parsing, concatenation and synthesis. Preprocessing of the input text in ISCII is done to expand abbreviations and numerals to their expanded text form, with the help of lookup tables. The parser parses the expanded text into a sequence of basic unit names. The parameters of the basic units are concatenated and after the modification by the activated knowledge, they are synthesized to give the speech output. The representation of the basic units are done using a speech production model in which the parameters Linear Prediction Coefficients (used for consonantal region) and formants and their bandwidths (used for vowel region) represent the vocal tract characteristics and the parameters pitch and gain represent the excitation characteristics. Various issues in constructing the basic units in this representation are the speech data collection, analysis for parameter extraction and coding of the unit using the parameters. The speech data was collected from a native Hindi speaker. The LPC, formants, pitch and gain parameters were extracted and the fields of the representation were filled in manually for each basic unit using the above parameters.

Chapter 4

COARTICULATION KNOWLEDGE FOR THE TEXT-TO-SPEECH SYSTEM

4.1 INTRODUCTION

This chapter discusses how to acquire, formulate and incorporate the coarticulation knowledge into the text-to-speech system for Hindi. Though coarticulation is an articulatory phenomenon, we study it in terms of acoustic features like formant transitions, durational changes and intensity variations so that it can be incorporated in a parameter based text-to-speech system. In section 4.2 we examine various coarticulation effects in Hindi speech and their acoustic manifestations. To construct a coarticulation knowledge base, we have to identify various patterns of acoustic manifestations and relate them to the phonetic context. Section 4.3 gives a classification scheme for various coarticulation patterns and formulates a fixed number of contextual rules. Section 4.4 discusses the representation and activation of the coarticulation knowledge in the text-to-speech system for Hindi.

4.2 NATURE OF COARTICULATION IN HINDI

We attempt to study and formulate the most common and obvious coarticulation effects as observed in Hindi speech. Coarticulation is more evident in the transitions between sound units than in the steady regions of sound units. The acoustic manifestations of the various coarticulation effects in Hindi speech are the following:

- (i) The transition ^{† १ ० १ ०} from C to V in a CV unit.
- (ii) The transition from V to C across a VC sequence.
- (iii) Transition from vowel to vowel in a VV sequence. These transitions are characterized by formant transition patterns and gain variation across the transition.
- (iv) Nasalization of the vowel. Nasalized vowels are characterized by the presence of antiformants in the spectrum.
- (v) Variation of the duration of the same unit in different contexts.

- (vi) Changes in the **spectral** features of a consonant in different contexts, namely the **CV**, **VC** and **CC** contexts.

Of these, the CV transition is embedded (or preserved) in the basic unit representation itself. The durational variation is discussed elsewhere (Rajesh Kumar, 1990). Our interest is primarily in VC, VV and CC transitions and nasalization of vowels. These are not represented or preserved in basic units since the context which decides the changes is not in the same unit but in two adjacent units. This means we need contextual rules to impart coarticulation across unit boundaries and also for nasalization of vowels.

In the study of the coarticulation, we have restricted the context to the immediately adjacent units. This greatly simplifies the study by reducing the number of cases to be considered. Coarticulation effects between units which are not immediately adjacent are less important for fluent speech production (O'Shaughnessy, 1987).

4 3 ACQUISITION AND FORMULATION OF COARTICULATION KNOWLEDGE FOR A TEXT-TO-SPEECH SYSTEM

The issues involved in acquisition and formulation of the coarticulation knowledge, from a point of incorporating into a text-to-speech system are:

- (i) Identification of the domain of coarticulation
- (ii) Classification of the coarticulation patterns
- (iii) Formulation of the coarticulation patterns in terms of suitable parameters

43.1 Identification of the domain of coarticulation

In general, the domain of coarticulation is the transition between basic units of speech. In particular, the transitions are between the neighbouring phonemes in a sequence of basic units. But, from an implementation point of view, all phoneme to phoneme transitions in a given text need not be considered since some of them are preserved in the basic units chosen for synthesis. The coarticulation information which are not captured by the basic units are to be brought in using rules. Table 4.1 shows the transitions to be considered for various types of basic units.

Table 4.1 Basic units and corresponding domain of **coarticulation**

Basic unit	Domain of coarticulation
Phoneme	CV, VC, CC and W
Diphone	W and CC
Characters	CC, W and VC
words	Some times possible between adjacent phonemes across word boundary
sentences	None

As seen from the table, if the basic units of synthesis are the phonemes, we have to incorporate all transitions between phonemes. Diphones capture the transition between phonemes. But diphones are cut in such a way that the **diphone** boundary coincides with the midpoint of a phoneme. Hence **diphone** junctions will be of the type CC and VV only. Characters capture all CV transitions. Words and sentences retain the coarticulation information. Since our basic speech units are the characters, we consider the junctions (transitions) between them. Characters can be a vowel (V), a consonant (C) and a consonant vowel sequence (CV). The transitions between these units can be only one-of the vowel to vowel (VV), vowel to consonant (VC) and consonant to consonant (CC). The coarticulation effect is brought about into these junctions by giving appropriate transition patterns of formants and also by adjusting the transition duration and the intensity.

43.2 Classification of various coarticulation patterns

Considering all combinations of any **two** basic units, the total number of junctions possible is about a few hundred. We classify these into a small number of basic transition patterns on the basis of the similarities in the articulatory transitions of the junctions concerned. The VC junctions (VC transitions) are grouped into distinct classes, each having a distinct place of articulation feature of the consonant C of the VC. The consonant C in the VC transition is the target to which the articulators move after the transition. Hence the nature of the transition pattern depends on the articulatory features of the consonant C (Ohman, 1966). We use the commonality in the articulatory features of C to group various VCs with same vowel and different consonants. One important feature of a consonant is its place of articulation. The place of articulation refers to the location of the constriction in the vocal tract. Another

important feature of the consonant articulation is the manner of articulation. The classification of Hindi consonants based on the features, place and manner of articulation, is given in Table 4.2. The classification of VC transition patterns is based on this.

Table 4.2 Classification of Hindi consonants using the place and manner of articulation features

Manner of articulation	Place of articulation							
		Bilabial	Dental	Alveolar	Retroflex	Palatal	Velar	Glottal
	Stop	p b ph bh	t d th dh		ʈ ɖ ʈh ɖh		k g kh gh	
	Affri.					c j ch jh		
	Nas.	m	n		ɳ	ɲ	ŋ	
	Fric.			s		ʃ		h
	Trill				r			
	Lat.			l				
	Glide		v			y		

Each major VC class corresponds to a place of articulation. Hence two VCs with same vowel quality but different consonants, both having the same place of articulation can be expected to have similar transition patterns of vowel formants. For example, the VC transitions /a:t/ (आत) and /a:th/ (आथ) have similar transition patterns. The minor variations within a VC class are attributed to the difference in the manner of articulation. **As** an example, consider the VC transitions /a:c/ (आच) and /a:y/ (आय). Though the patterns are similar, there are differences in the extent of formant changes, transition duration and the gain variation across the transition. In order to account for this, each VC class is again divided into subgroups on the basis of manner of articulation of the consonant. Thus each nonempty cell in Table 4.2 has a corresponding VC subgroup consisting of transitions from various vowels to the consonants of that cell. Each VC subgroup is further divided using a third attribute, the vowel quality. In Hindi, there are five different vowels. Each VC subgroup contains a formant transition pattern for each of the vowels. The validity of the above classification of VC transitions is that it is based on the articulatory similarity of the units involved.

Vowel to vowel transitions (VV) in Hindi are only a few of the various VV combinations (Since there are five pure vowels in Hindi, there is a total of 25 distinct VV sequences). The transition is characterized by the gradual formant transition from the first vowel to the next. The perceptual effect of this is the presence of a glide, decided by the vowels involved in the transition.

The coarticulation across consonant to consonant (CC) transition is complicated and involves changes in spectral features of the consonant. The current text-to-speech system does not provide the flexibility for consonantal spectral manipulation. We have not attempted to study coarticulation in CC sequences. Instead, some common and perceptually significant effects associated with CC sequences in Hindi, which involve only duration and gain manipulation, are formulated, namely: (i) Release of word final cluster, (ii) Gemination rule and (iii) Lengthening of the consonant before a glide (all are explained latter).

4.33 Formulation of coarticulation patterns as rules

Based on the classification of the VC transition patterns, we formulate a set of basic transition patterns. For each VC subgroup (all VCs formed from a vowel and the consonants having the same place and manner of articulation feature) a transition pattern is formulated. These basic transition patterns are referred to as rules here after. Each rule specifies the formant frequency transition pattern and the transition duration. The parameters required for the rule specification are obtained from the analysis of natural speech data. A systematic data collection and analysis strategy are followed in which nonsense words containing these junctions, uttered by a speaker are collected and analyzed and the transition pattern is formulated in terms of formant frequency changes and the duration of transition.

Besides the VC transition rules, we have formulated rules for vowel to vowel transition, nasalization of vowels and cluster consonants. The components of the rule set are the following:

VC transition rules	70 Nos
Nasalization rules	2 “
Vowel to vowel transition rule	1 “
Cluster consonant rules	3 “

4.3.3.1 Vowel to consonant transition rules

According to the classification of VC transitions, we have the following major classes of transition patterns:

1. Velar
2. Palatal
3. Retroflex
4. Alveolar
5. Dental
6. Bilabial

Each of these is further divided into subgroups by manner of articulation feature. These can be:

1. Stop
2. Affricate
3. Nasal
4. Fricative
5. Trill
6. Lateral
7. Glide

Rules formulated for these classes of transitions basically specify the formant transition patterns. The parameters used are the formant frequency changes for the first three formants and the duration of transition. The formants either fall or rise from their steady values. The change is expressed as the percentage of the steady state value (Negative for falling and positive for rising). Each subgroup contains a transition pattern for each of the five different vowel qualities. The numerical values used in the rules are obtained from actual speech data. In the following, first we give data collection and analysis details required for the rule formulation and then we list the rules as they are formulated.

433.1.1 Data collection and analysis for rule formulation

The parameter values which specify the VC transition namely the formant frequency changes for the first three formants and the transition duration are obtained

by analysing the VC transitions in the natural speech data. We use isolated utterances of the VC sequences as the speech data rather than the VC sequences in continuous speech. This is because we consider the coarticulation between two adjacent phonemes only. Spectrograms of isolated VC utterances are obtained using the KAY DSP Sonagraph. From these the formant frequency shifts and the transition duration of the VC transition are obtained. The formant shifts are expressed as the percentage of corresponding steady formant values of the vowel.

The patterns formulated are listed for each VC subgroup.

4.33.1.2 Velar/Glottal VC transitions

The subgroups under the velar VC transitions are those of stops and fricatives.

(i) Transitions involving velar stops

These VC transitions involve one of the following consonants: /k/ (क), /g/ (ग), /kh/ (ख) and /gh/ (घ). The spectrograms of VC sequences involving the consonant /k/ (क) with different vowels are analyzed to formulate the formant transition patterns. These are listed in Table 4.3 for different vowels of Hindi. The transition is specified as the percentage changes in steady formant values of the vowel and the transition duration in frames of 6.4 msec. Fig. 4.1 shows the spectrogram of a sample VC transition for the VC /a:k/ (आक). The transition patterns given in the table are shown diagrammatically in Fig. 4.2. The similarity of the formant transition pattern in the spectrogram and the transition pattern for /a:k/ (आक) in Fig. 4.2 may be noted.

Table 4.3 Table of VC formant transitions from different vowels to velar stops. The transition is specified as the percentage changes in steady formant values of the vowel and the transition duration in frames of 6.4 msec.

Vowel	a	e	i	o	u
% change in F ₁	-40	-9	-14	-30	-30
% change in F ₂	0	0	-13	0	0
% change in F ₃	0	-15	-14	0	0
Trans. Dur. in frames	4	5	4	4	4

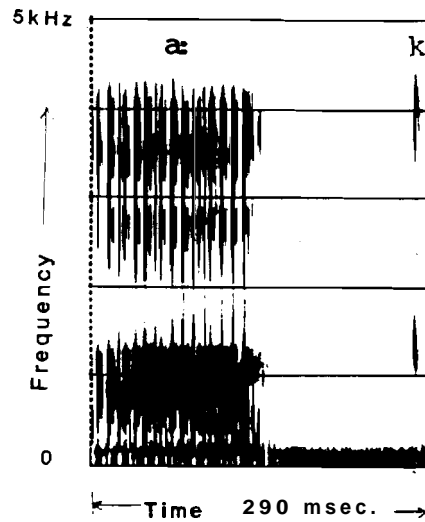


Fig. 4.1 Spectrogram of VC transition /a:k/ (आक)

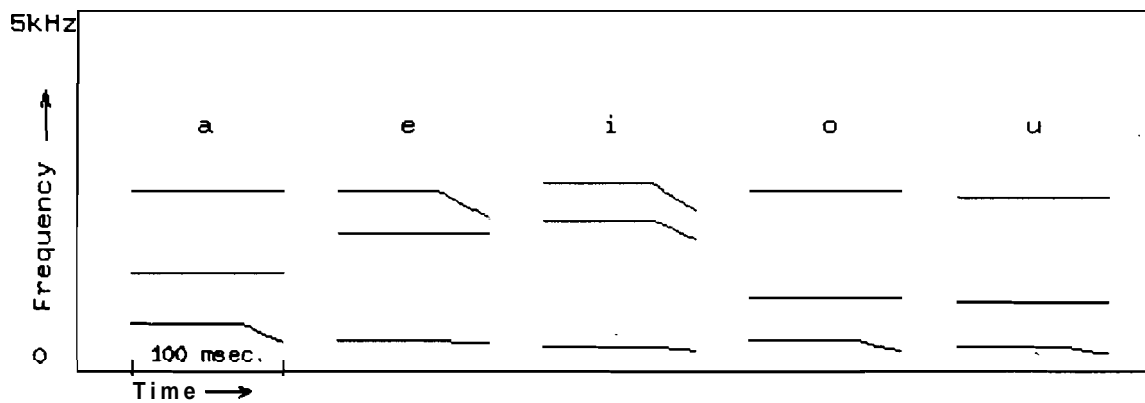


Fig. 4.2 VC formant transition in the context of velar stops. These patterns are generated from the specification in Table 4.3. The first one of these corresponds to the spectrogram in Fig. 4.1

(ii) Transitions involving the glottal fricative

The VC transitions where C is the velar fricative consonant /h/ (ह) is not formulated because it is seen that /h/ (ह) has got the same formant structure of the preceding vowel in a VC (/h/ being a glottal fricative, does not require any constriction of the vocal tract during its articulation).

In the above VC transitions involving velar consonants, we have not distinguished between voiced and unvoiced consonants. But in reality, the VC transitions involving the unvoiced consonants are shorter in duration than those involving voiced consonants. This is mainly due to the abrupt termination of the vowel prior to the silence region of the unvoiced stop. This prevents the formants from reaching their target values. This effect is brought into the transition by abruptly reducing the vowel gain in the end of the VC transition at the time of synthesis.

433.13 Palatal VC transitions

The palatal consonants in Hindi are divided into the affricates, fricative and glide based on the manner of articulation.

(i) Transitions involving affricates

The palatal affricate consonants are /c/ (च), /ch/ (छ), /j/ (ज) and /jh/ (झ). The characteristic manner of articulation of these consonants is the initial silence or voicing followed by a fricative turbulent air flow giving rise to the production of the **frication** noise. The transition patterns formulated from the spectrograms of VC sequences involving /c/ (च) are given in Table 4.4. These are shown diagrammatically in Fig. 4.3. Fig. 4.4 shows a sample spectrogram for the VC /a:c/ (आच).

Table.4.4 VC transitions in the context of palatal affricates

Vowel	a	e	i	o	u
% change in F ₁	-33	-30	0	-27	-20
% change in F ₂	30	15	0	70	70
% change in F ₃	-10	10	0	-15	-10
Trans. Dur. in frames	6	5	1	7	7

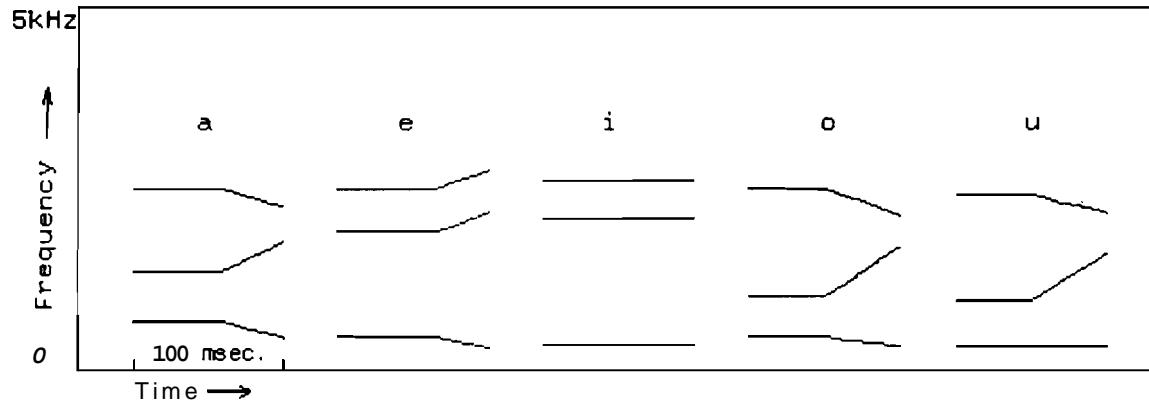


Fig. 4.3 VC transitions in the context of palatal affricates These are generated from Table 4.4.

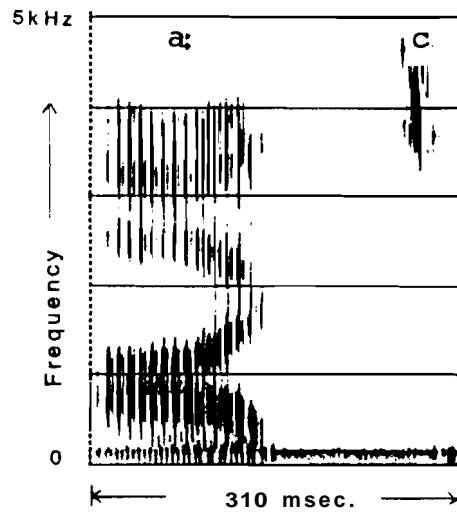


Fig. 4.4 Spectrogram of VC transition /a:c/ (आच)

Table. 4.4 VC transitions in the context of palatal affricates

Vowel	a	e	i	o	u
% change in F ₁	-33	-30	0	-27	-20
% change in F ₂	30	15	0	70	70
% change in F ₃	-10	10	0	-15	-10
Trans. Dur. in frames	6	5	1	7	7

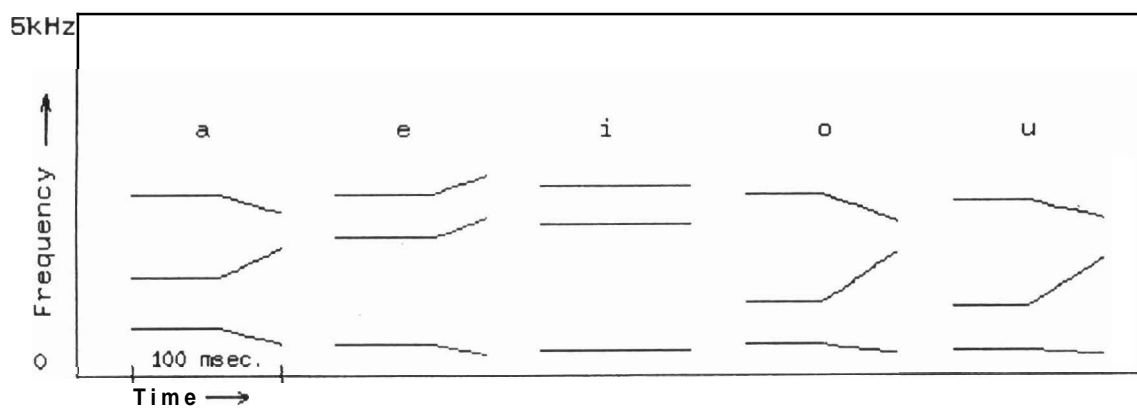


Fig. 4.3 VC transitions in the context of palatal affricates These are generated from Table 4.4.

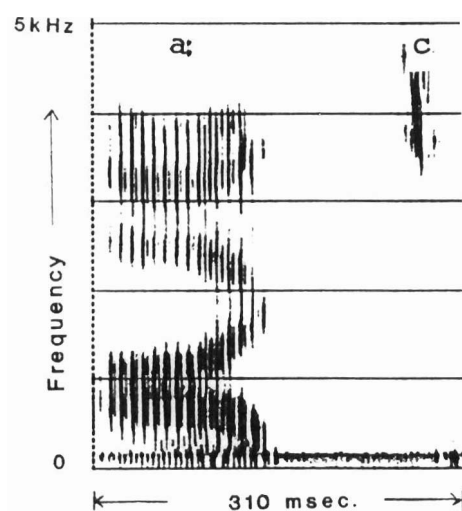


Fig. 4.4 Spectrogram of VC transition /a:c/ (आच)

(ii) **Transitions** involving the palatal fricative

The formant transitions to the palatal fricative /ʃ/ (श) is formulated below (Table 4.5 and Fig. 4.5) for all vowels as observed from the analysed data. A sample spectrogram is given in Fig.4.6 for /a:ʃ/ (आश).

Table 4.5 VC transitions in the context of palatal fricative

Vowel	a	e	i	o	u
% change in F ₁	-12	0	0	0	14
% change in F ₂	10	0	0	32	61
% change in F ₃	-21	0	0	-5	-7
Trans. Dur. in frames	6	1	1	7	6

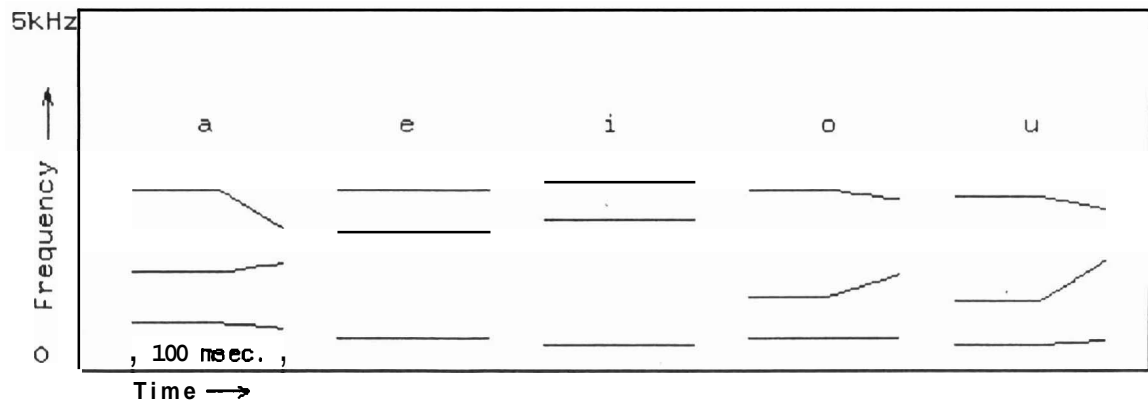


Fig. 4.5 VC transitions in the context of palatal fricative generated from Table 4.5

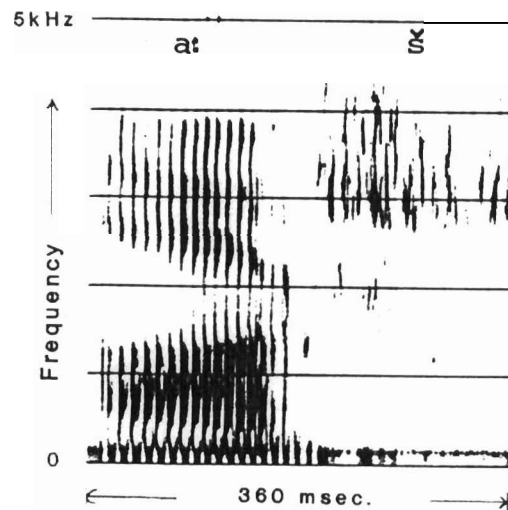


Fig. 4.6 Spectrogram of /a:ʃ/ (आश)

(ii) **Transitions involving the palatal fricative**

The formant transitions to the palatal fricative (झ) is formulated below (Table 4.5 and Fig. 4.5) for all vowels as observed from the analysed data. A sample spectrogram is given in Fig.4.6 for /a:ʒ/ (आझ).

Table 4.5 VC transitions in the context of palatal fricative

Vowel	a	e	i	o	u
% change in F ₁	-12	0	0	0	14
% change in F ₂	10	0	0	32	61
% change in F ₃	-21	0	0	-5	-7
Trans. Dur. in frames	6	1	1	7	6

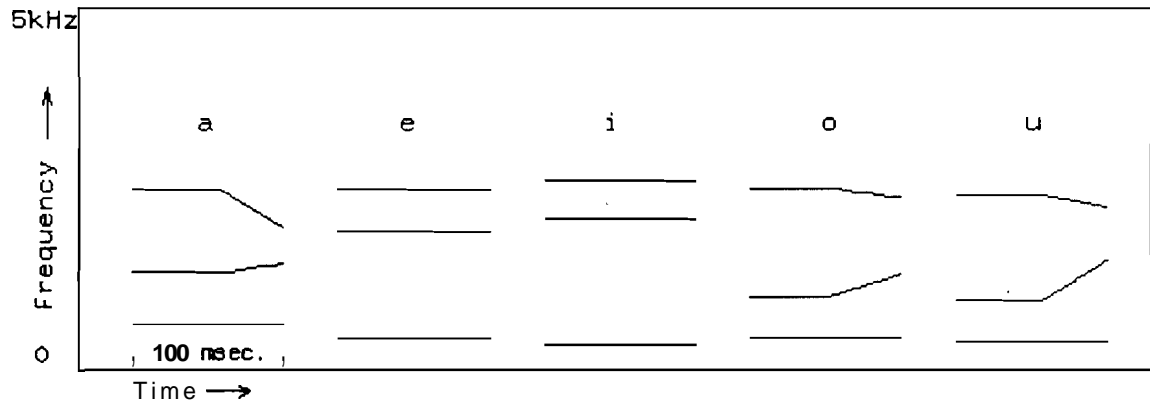


Fig. 4.5 VC transitions in the context of palatal fricative generated from Table 4.5

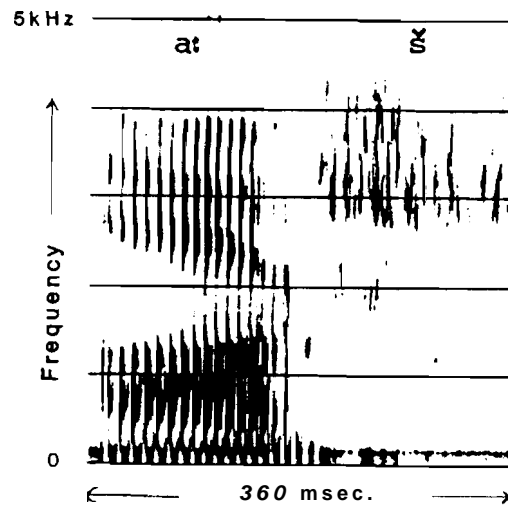


Fig. 4.6 Spectrogram of /a:ʒ/ (आझ)

(iii) Transitions involving the palatal glide

The VC transitions involving the palatal glide /y/ (य) are shown in Fig. 4.7. Fig. 4.8 shows the sample spectrogram of the VC sequence /a:y/ (आय). The vowel to glide transitions are actually not formulated. They are obtained by the interpolation of formants of the vowel and the glide.

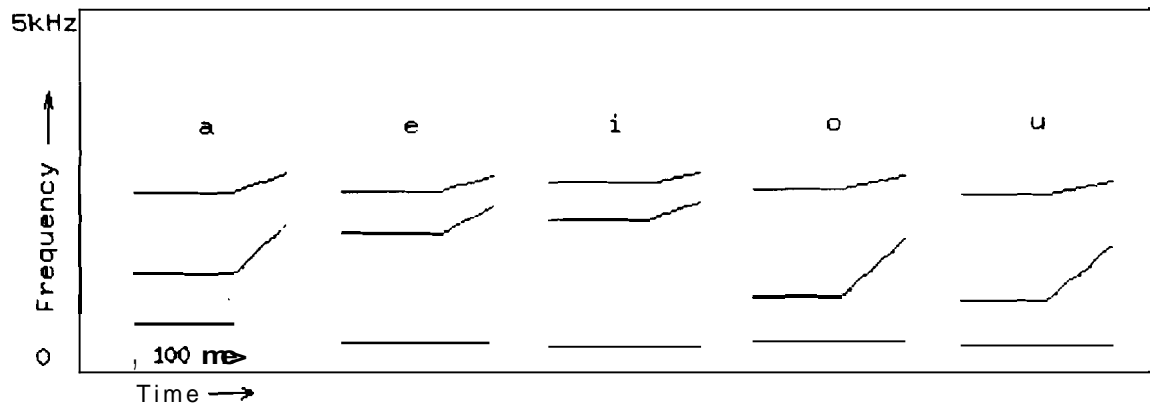


Fig. 4.7 VC transitions in the context of palatal glide

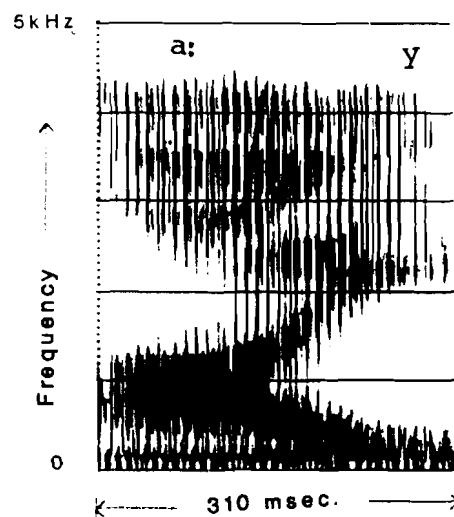


Fig. 4.8 Spectrogram of /a:y/ (आय)

433.1.4 Retroflex VC transitions

The retroflex category of VC transitions are those involving stops, fricative and nasal.

(iii) Transitions involving the palatal glide

The VC transitions involving the palatal glide /y/ (ꣳ) are shown in Fig. 4.7. Fig. 4.8 shows the sample spectrogram of the VC sequence /a:y/ (ꣳꣳ). The vowel to glide transitions are actually not formulated. They are obtained by the interpolation of formants of the vowel and the glide.

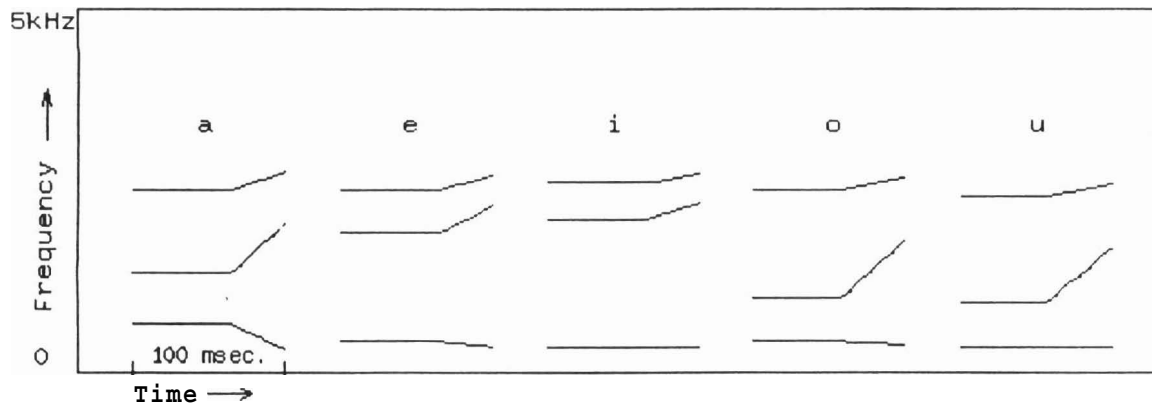


Fig. 4.7 VC transitions in the context of palatal glide

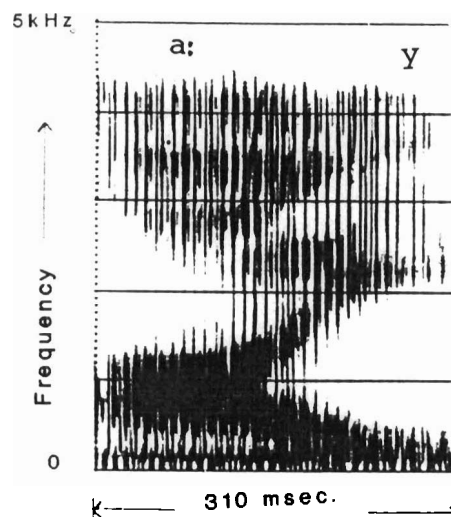


Fig. 4.8 Spectrogram of /a:y/ (ꣳꣳ)

433.1.4 Retroflex VC transitions

The retroflex category of VC transitions are those involving stops, fricative and nasal.

(i) **Transitions involving retroflex stops**

The VC transitions in the context of the retroflex stops /t/ (ट), /tN (᳚), /d/ (ड) and /dh/ (᳚) for different vowels are given in Table 4.6 and illustrated in Fig. 4.9. Fig. 4.10 shows the spectrogram of the VC sequence /a:t/ (अट).

Table 4.6 VC transitions in the context of retroflex stops

Vowel	a	e	i	o	u
% change in F ₁	-40	0	0	0	-13
% change in F ₂	20	0	0	25	25
% change in F ₃	-35	-10	-8	-35	-38
Trans. Dur. in frames	4	4	5	4	5

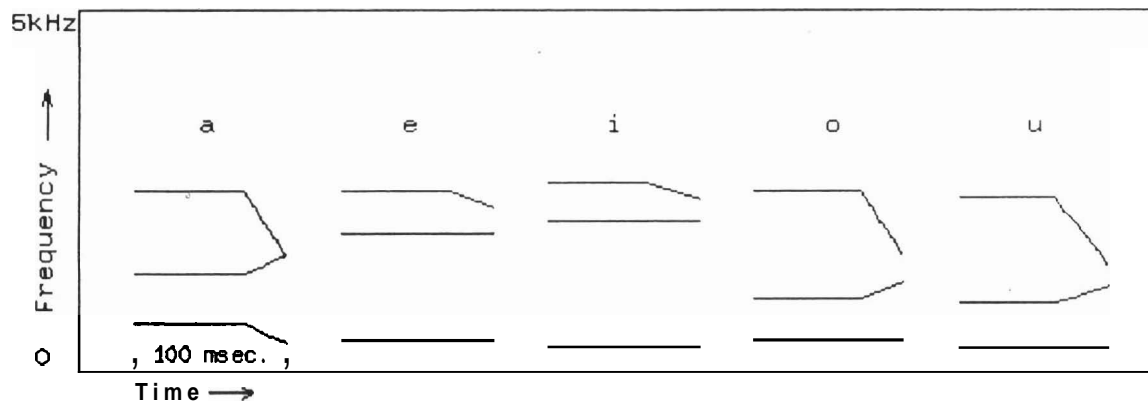


Fig.4.9 VC transitions in the context of retroflex stops generated from Table 4.6

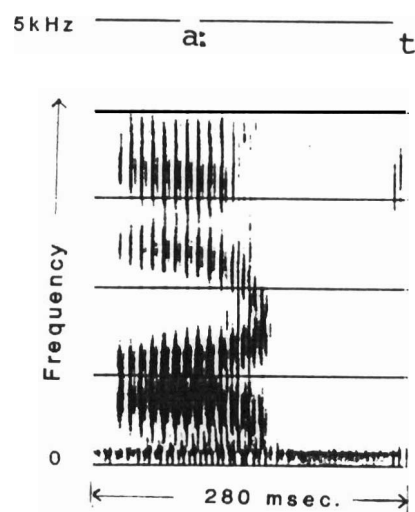


Fig.4.10 Spectrogram of /a:t/ (अट)

(i) **Transitions involving retroflex stops**

The VC transitions in the context of the retroflex stops /t/ (ṭ), /tʰ/ (ṭʰ), /d/ (ḍ) and /dʰ/ (ḍʰ) for different vowels are given in Table 4.6 and illustrated in Fig. 4.9. Fig. 4.10 shows the spectrogram of the VC sequence /a:t/ (आṭ).

Table 4.6 VC transitions in the context of retroflex stops

Vowel	a	e	i	o	u
% change in F ₁	-40	0	0	0	-13
% change in F ₂	20	0	0	25	25
% change in F ₃	-35	-10	-8	-35	-38
Trans. Dur. in frames	4	4	5	4	5

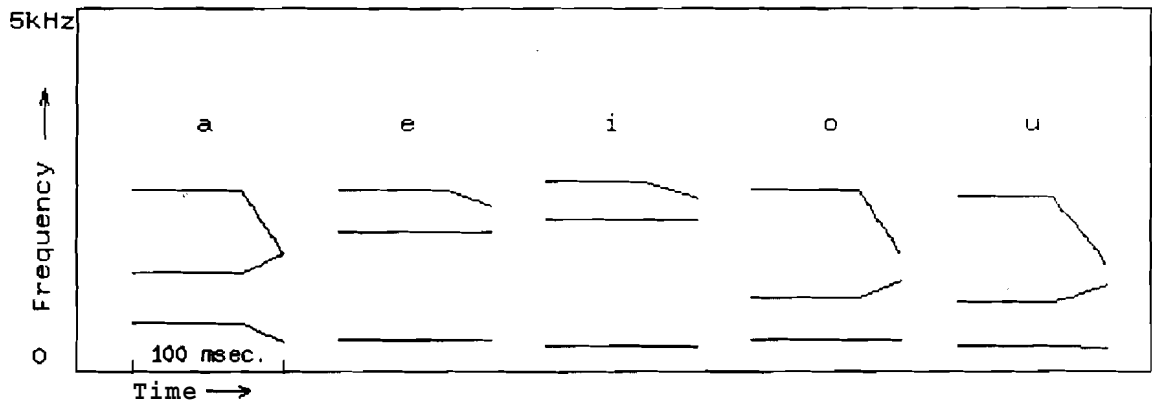


Fig.4.9 VC transitions in the context of retroflex stops generated from Table 4.6

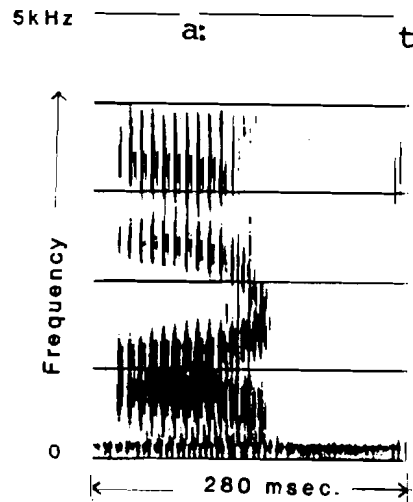


Fig.4.10 Spectrogram of /a:t/ (आṭ)

(ii) Transitions involving retroflex fricative /ʂ/ (ष)

Phonologically there is no retroflex fricative consonant in Hindi. It has merged with the palatal. However, in writing system they are different. We use the same VC transitions for the palatal and retroflex fricatives.

(iii) Transitions involving the retroflex nasal

The VC formant transitions to the retroflex nasal /ɳ/ (ण) are formulated in Table 4.7 and are illustrated in Fig. 4.11 along with a sample spectrogram of /a:ɳ/ (आण) in Fig. 4.12. These patterns are similar to those of retroflex stops but the transition duration is more. Besides, the vowel is nasalized in this case.

Table 4.7 VC transitions in the context of retroflex nasal

Vowd	a	e	i	o	u
% change in F ₁	-26	0	0	-11	-13
% change in F ₂	18	-12	-15	35	50
% change in F ₃	-35	-2	8	-40	-35
Trans. Dur. in frames	8	8	7	9	9

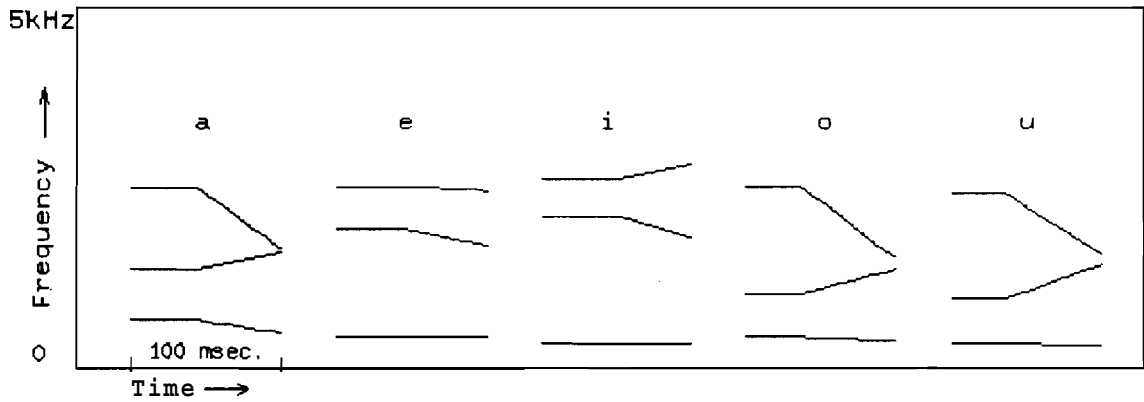


Fig. 4.11 VC transitions in the context of retroflex nasal generated from Table 4.7

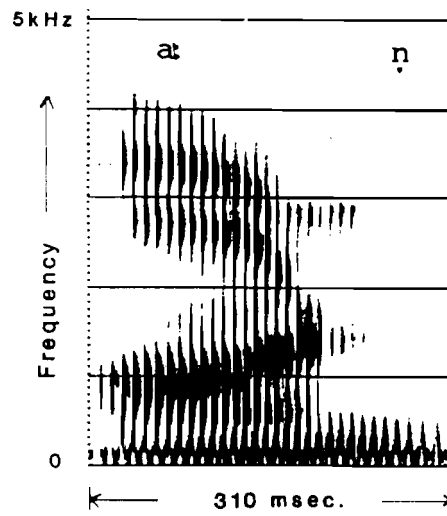


Fig. 4.12 Spectrogram of /a:ŋ/ (आण्)

433.15 Alveolar VC transitions

The alveolar **VC** transitions are subgrouped based on the manner of articulation into those involving fricatives, trill and lateral. The formulation of the **VC** transition patterns to each of these is shown in Tables 4.8 through 4.10 along with illustrations and sample spectrograms (Figs. 4.13 - 4.18).

(i) **Transitions** involving alveolar fricative /s/ (स)

Table 4.8 **VC** transitions in the context of alveolar fricative

Vowel	a	e	i	o	u
% change in F ₁	-12	0	0	0	0
% change in F ₂	10	-22	-18	43	40
% change in F ₃	0	-3	-14	0	0
Trans. Dur. in frames	8	8	8	8	8

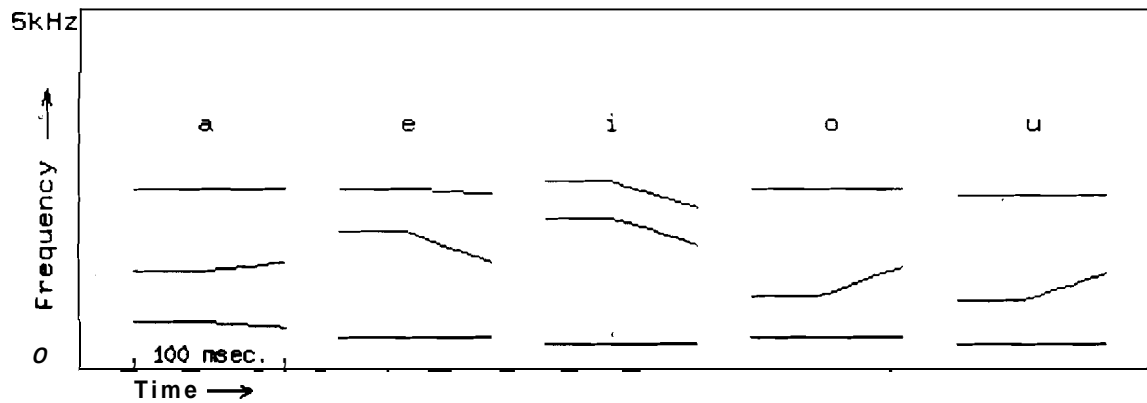


Fig. 4.13 **VC** transitions in the context of alveolar fricative generated from Table 4.8

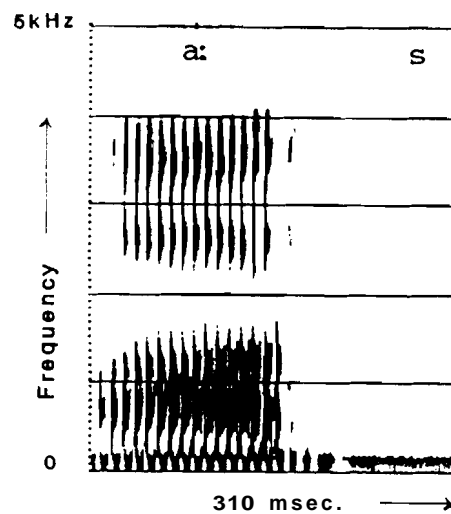


Fig. 4.14 Spectrogram of /a:s/ (आस)

(ii) Transitions involving lateral /l/ (ਲ)

Table 4.9 VC transitions in the context of the lateral

Vowel	a	e	i	o	u
% change in F ₁	-19	0	0	0	0
% change in F ₂	27	-17	-23	50	39
% change in F ₃	0	0	-10	0	0
Trans. Dur. in frames	7	9	7	8	8

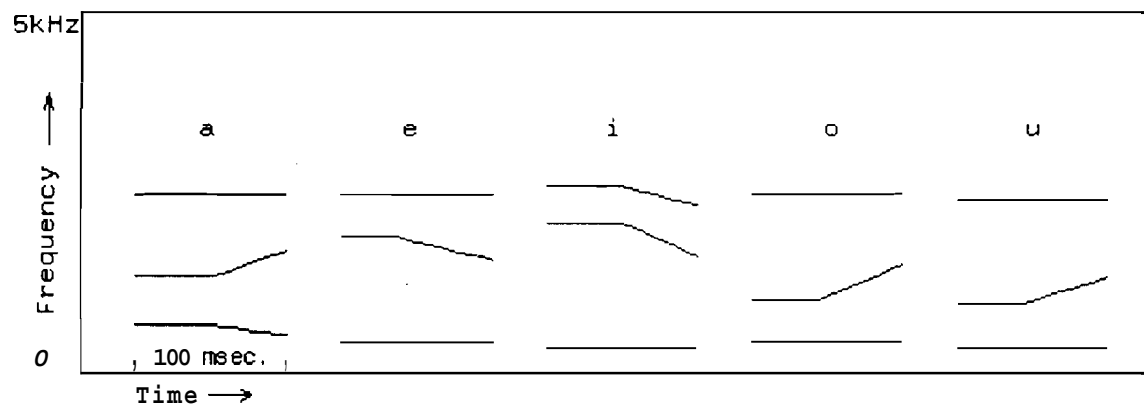


Fig. 4.15 VC transition in the context of the lateral generated from Table 4.9

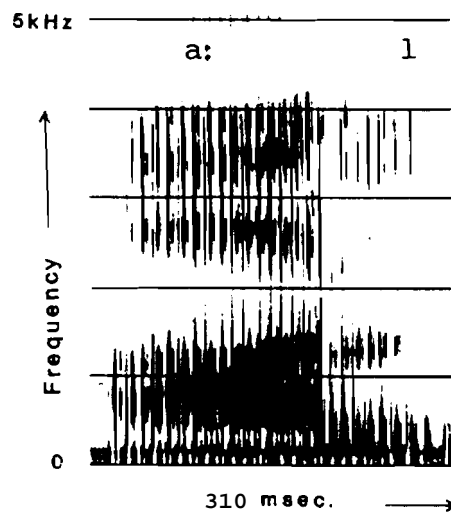


Fig. 4.16 Spectrogram of /a:l/ (ਅਲ)

(iii) **Transitions involving trill /r/ (ठ)**

Table 4.10 VC transitions in the context of the trill

Vowel	a	e	i	o	u
% change in F ₁	-6	9	29	-9	11
% change in F ₂	7	-22	-26	23	30
% change in F ₃	-20	0	-14	-26	-32
Trans. Dur. in frames	6	8	9	8	6

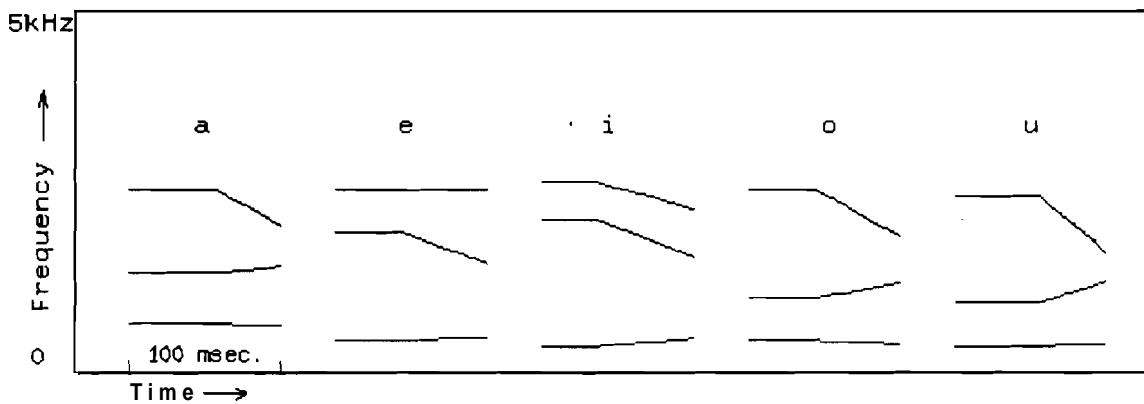


Fig. 4.17 VC transitions in the context of the trill generated from Table 4.10

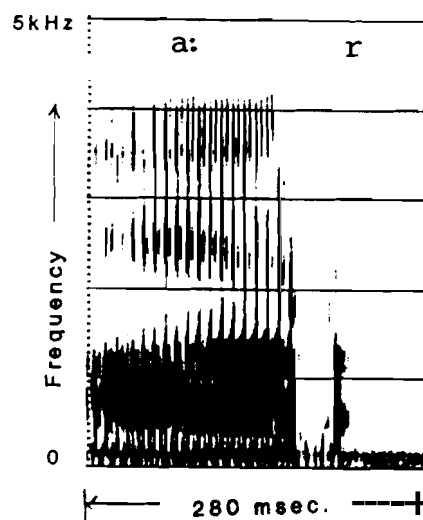


Fig. 4.18 Spectrogram of /a:r/ (आऱ)

433.1.6 Dental VC transitions

The dental consonants are divided into stops, nasal and glide.

(i) **Transitions involving dental stops**

The transitions to dental stops /t/ (त), /th/ (थ), /d/ (द) and /dh/ (ध) are given in Table 4.11 for various vowels. The transitions are illustrated in Fig. 4.19 and a sample spectrogram of the VC/a:t/ (आत) is shown in Fig. 4.20.

Table 4.11 VC transitions in the context of dental stops

Vowel	a	e	i	o	u
% change in F ₁	-30	0	0	0	0
% change in F ₂	25	-15	-18	36	33
% change in F ₃	0	0	-14	0	0
Trans. Dur. in frames	4	6	5	4	5

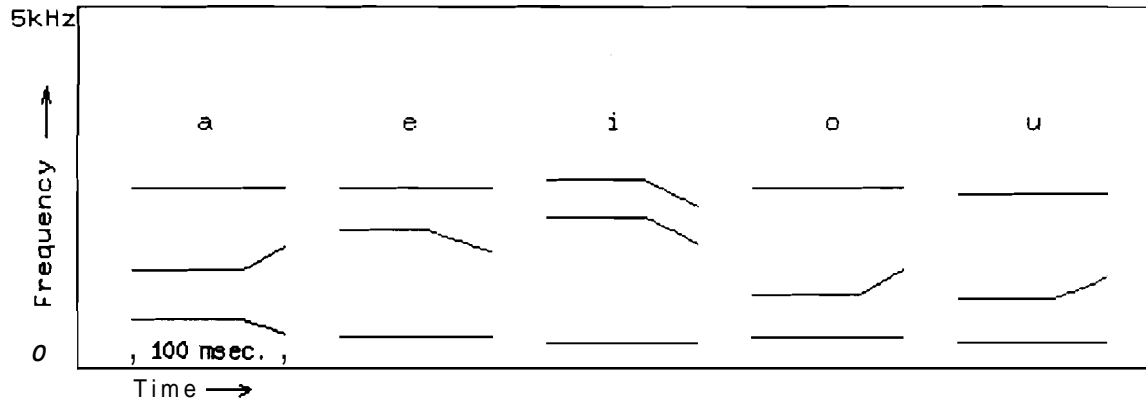


Fig. 4.19 VC transitions in the context of dental stops generated from Table 4.11

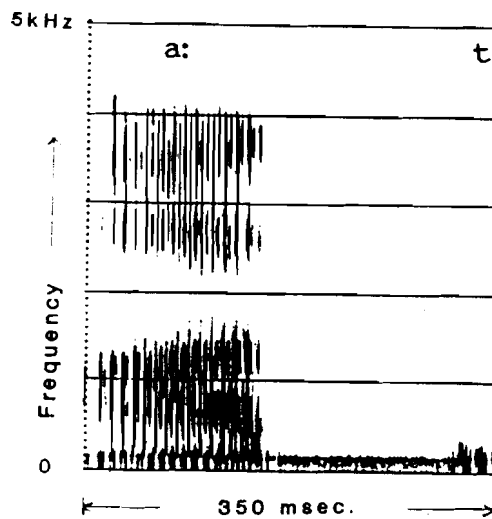


Fig. 4.20 Spectrogram of /a:t/ (आत)

(ii) **Transitions involving the nasal**

The VC transition patterns of various vowels into dental nasal /n/ (न) are as formulated in the Table 4.12 and are illustrated in Fig. 4.21 along with the spectrogram of /a:n/ (अन) in Fig. 4.22.

Table 4.12 VC transitions in the context of dental nasal

Vowel	a	e	i	o	u
% change in F ₁	-16	0	0	0	0
% change in F ₂	21	-15	-22	63	44
% change in F ₃	0	0	-15	0	0
Trans. Dur. in frames	6	7	6	8	5

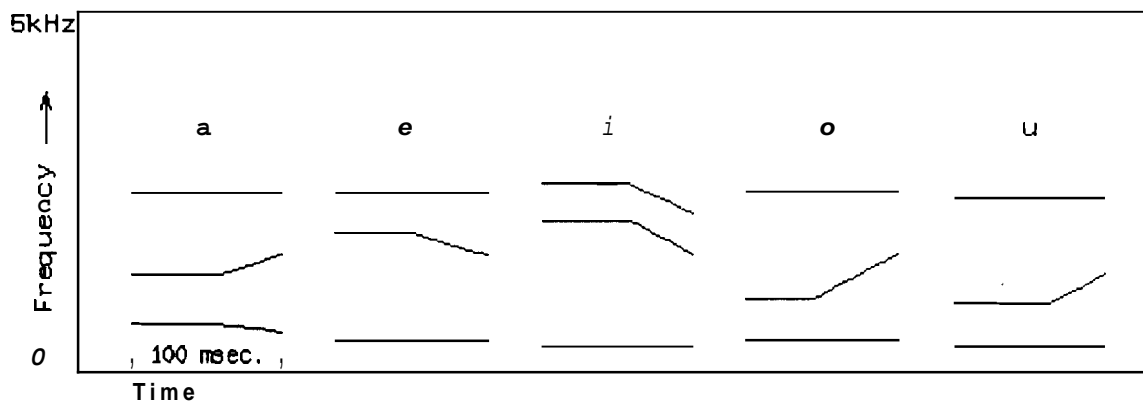


Fig. 4.21 VC transitions in the context of dental nasal generated from Table 4.12

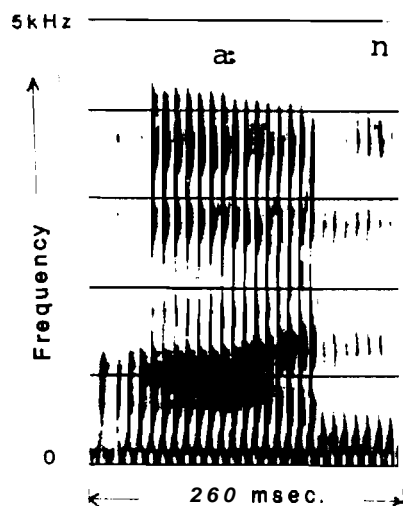


Fig. 4.22 Spectrogram of /a:n/ (अन)

(iii) Transitions involving the glide

The labio-dental glide /v/ (व) has got a vowel like formant structure. The transition patterns from the vowels to this are given in Fig. 4.23. Fig. 4.24 shows the sample spectrogram for /a:v/ (वा). As in the case of palatal glides, the dental glide transitions are generated by interpolating the vowel formants with those of the glide.

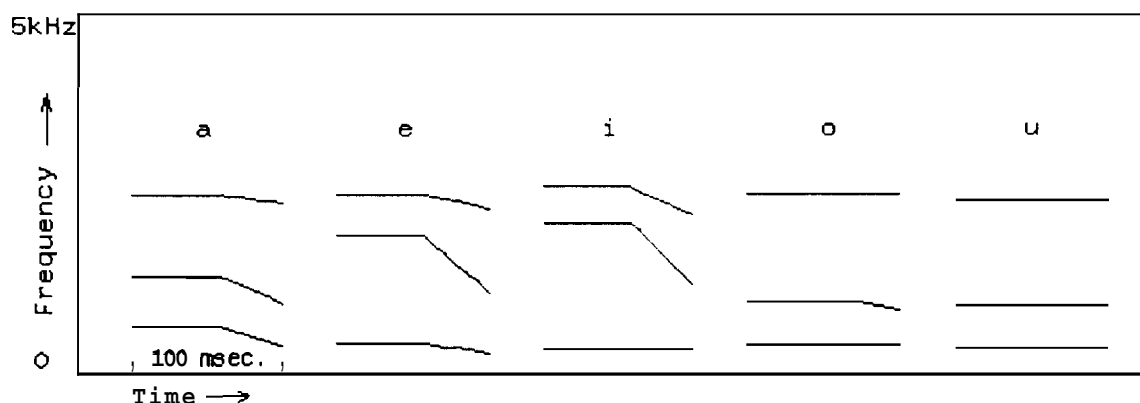


Fig. 4.23 VC transitions in the context of labio-dental glide

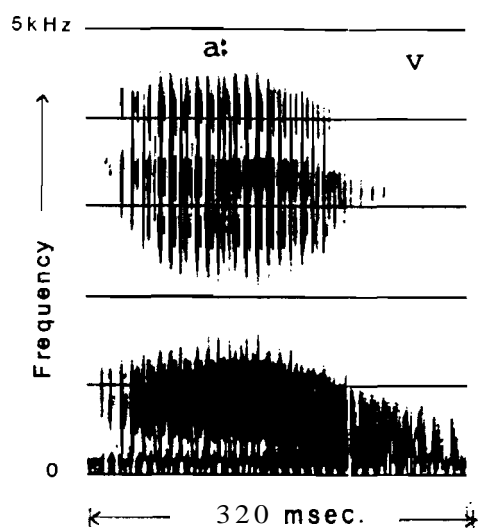


Fig. 4.24 Spectrogram of /a:v/ (वा)

433.1.7 Bilabial VC Transitions

The bilabial transitions are those involving the bilabial stops and the bilabial nasal.

(i) **Transitions involving stops**

The vowel formant transitions to bilabial stops /p/ (प), /ph/ (फ), /b/ (ब) and /bh/ (भ) are as given in Table 4.13. These are illustrated in Fig. 4.25 and the sample spectrogram of the VC sequence /a:p/ (आप) in Fig. 4.26. Note that for bilabials none of the formants show a rising transition.

Table 4.13 VC transitions in the context of the bilabial stops

Vowel	a	e	i	o	u
% change in F ₁	-40	-20	0	0	0
% change in F ₂	-30	-38	-32	-20	-26
% change in F ₃	0	-8	-20	0	0
Trans. Dur. in frames	4	4	3	4	4

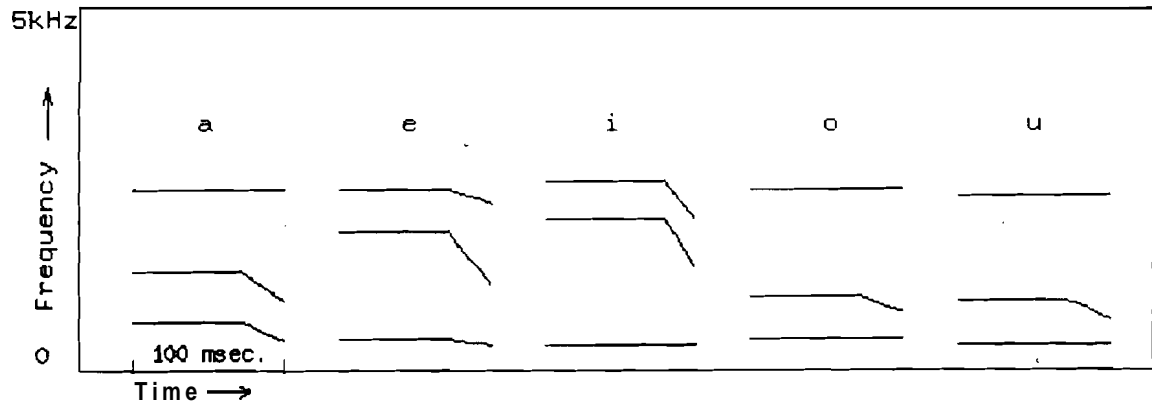


Fig. 4.25 VC transitions in the context of bilabial stops generated from Table 4.13

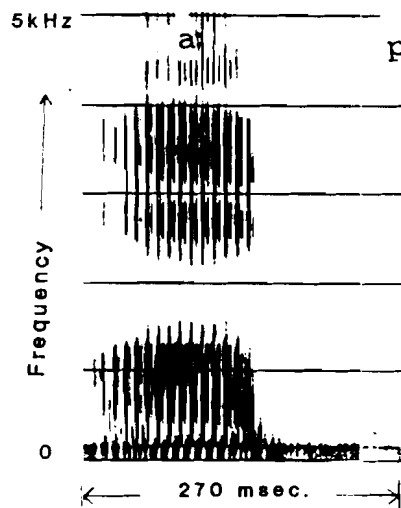


Fig.4.26 Spectrogram of the VC /a:p/ (आप)

(ii) Transitions involving nasal

The VC transition patterns for the bilabial nasal /m/ (म) are given in Table 4.14 and illustrated in Fig. 4.27. The sample spectrogram for the VC transition /a:m/ (आम) is shown in Fig. 4.28

Table 4.14 VC transitions in the context of bilabial nasal

Vowel	a	e	i	o	u
% change in F ₁	-38	0	0	-20	0
% change in F ₂	-25	-35	-45	-10	0
% change in F ₃	0	-15	-15	0	0
Trans. Dur. in frames	4	6	6	5	1

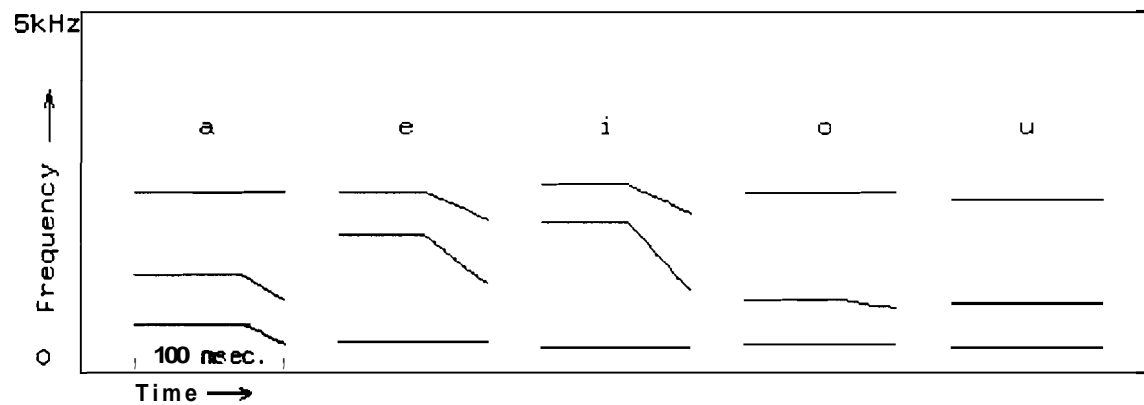


Fig. 4.27 VC transitions in the context of bilabial nasal generated from Table 4.14

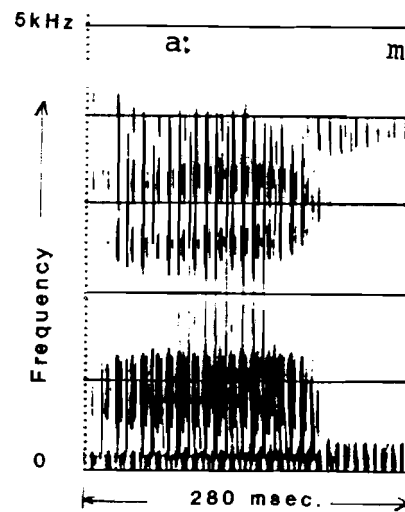


Fig. 4.28 Spectrogram of /a:m/ (आम)

The above six categories covered all VC transitions in Hindi. In the transitions involving stop consonants, we have to distinguish between those involving unvoiced stops from those for voiced ones. In both cases the transition pattern is same and is as given in the tables but transition duration will be less for unvoiced stops due to the sudden termination of vowels. This difference is taken into account at the time of synthesis by abruptly reducing the vowel gain in the final one or two frames of the VC transition.

4 3 3 2 . Nasalization rules

The articulation of nasal consonants involves the excitation of the nasal cavity besides the vocal tract. But because of physiological constraints, the opening and closing of the nasal cavity is a gradual process. This causes some amount of nasalization of any surrounding vowel. The perceptual effect of the nasalization is very significant in **recognizing** the nasal consonant. The vowel nasalization is important in the perception of nasal consonants in a formant synthesis system. Two nasalization rules - one for CV and other for VC context - simulate the nasalization effect. It is to be noted that vowel nasalization is different from the transitions (CV and VC) involving nasals and hence is not automatically taken care of by the latter.

The nasals and nasalized vowels are characterized by the presence of antiresonances in the frequency spectrum. In a formant synthesis system, the antiresonance is simulated using a pole-zero pair, before the first formant of the vowel. The activation of the nasalization rules is explained in the section on knowledge activation.

4 3 3 3 Vowel to vowel transition rules

In Hindi, there are words in which two vowels occur in a sequence. Examples are /kai:/ (कई), /hua:/ (हुआ), /bhai:/ (भाई) etc. When we synthesize such words, sequences of type **W** occur between basic units. It is necessary to give the vowel to vowel transition pattern since mere concatenation of vowels does not preserve the naturalness in a V to V transition. It is important to note that in vowel to vowel transitions, a glide (semivowel) decided by the vowels involved is perceived. For example, /kai:/ (कई) is perceived as /kayi:/ (कयी), /hua:/ (हुआ) is perceived as /huva:/ (हुवा) etc. This means that besides giving appropriate transition, we have to

increase the duration of transition in order to accommodate the glide. VV transitions are characterized by gradual transition of formants from one vowel to the following one. It is observed that this formant transition is almost equivalent to the one obtained by the interpolation of corresponding formants of the vowels involved. This allows us to take care of all vowel to vowel transitions by a single rule which does the interpolation of vowel formants.

433.4 Cluster consonant rules

We have formulated a few rules based on some common and simple effects observed in Hindi consonant clusters. These rules manipulate the duration and the release part (see below) only of the clusters and do not modify the spectral information of the constituent consonants. These rules are to account for the following effects:

- (i) Releasing of the cluster consonant in the word final position
- (ii) Gemination
- (iii) Lengthening of consonants before glides

(i) Releasing of word final cluster

Releasing refers to the small vowel-like segment which follows the cluster consonant in the word final position. In other positions, a cluster is usually associated with a vowel and hence the release effect does not apply to them. Examples of words containing word final cluster are /anth/ (अंत), /patr/ (पत्र), /sama:pt/ (समाप्त) etc. The release is very important in the perception of the final consonant in the word final cluster. For example, among the words listed above, the final consonant /t/ (त) in /sama:pt/ (समाप्त) is seen to be difficult to identify by listening when the word is **synthesized** without the word final release. The release segment is implemented using formants.

(ii) Gemination rule

A geminated cluster is the one which contains two consonants of the same place of articulation. Some examples of words containing these clusters are /bacca:/ (बच्चा), /chabbi:s/ (छब्बीस), /kutta:/ (कुत्ता), /billi:/ (बिल्ली) etc. The gemination rule concatenates two stop consonants to form a cluster as follows: (i) It suppresses the burst, if any, of the first consonant in the cluster (ii) It combines the silence (voicing for voiced consonants) of both consonants in the cluster to get a long silence (or voiced) region.

The above two are required for the proper perception of the geminated cluster. The geminated clusters of other consonants like nasals, lateral etc. are generated by combining the consonant regions which is found to be satisfactory.

(iii) Lengthening of consonants before glides

It is observed that the first consonant in consonant clusters in which second is a glide (/y/ (य) and /v/ (व)) is longer in duration than when they are in a CV unit. This effect is found to be significant perceptually. For example, the word /vidhya:rthi/ (विद्यार्थि) synthesized without the lengthening of /dh/ (ध) in the cluster /dhya:/ (द्या) is perceived to be less natural. From the analysis of some clusters in natural speech, it is seen that the lengthening of the consonant is about 1.5 times its duration in a CV context. The lengthening is done in the initial region (silence or voicing) of the consonant.

4.4 INCORPORATION OF THE COARTICULATION KNOWLEDGE TO THE TEXT-TO-SPEECH SYSTEM

The coarticulation effects as formulated in the preceding sections are a collection of rules which specify the transition patterns and other contextual parameter modifications. Incorporation of the coarticulation into a text-to-speech system involves the application of these rules to modify the default parameters of the basic units in the input text before synthesizing them, by making use of the phonetic context of the basic units in the text. The issues to be addressed in this include use of appropriate representation and activation methods for the coarticulation rules. Depending on the properties of the knowledge, we have to select an appropriate knowledge representation method **from** various options like tables, production systems, semantic networks etc. The activation method should make use of appropriate search methods for matching the context and heuristics for minimizing the search space and time. The representation and activation scheme is constrained by the fact that the activation process is **very** much dependent on the synthesis process. The knowledge representation and activation is guided by (or closely tied to) the synthesis module. This is because of the fact that coarticulation modifies more than one parameter of the basic unit and should be applied before it is synthesized. In the following sections, first we consider the architecture of the synthesis module which should be made clear before the

discussion of the **coarticulation** module. Then the knowledge representation and activation issues are discussed.

4.4.1 Architecture of the synthesis module

We can distinguish between two different approaches of organizing the synthesis process: (i) Synthesizing directly from the representation of units. The rules are activated as and when they are needed in the synthesis process. This approach does not require any buffers as there is no storage of the parameter contours for a full sentence or clause. (ii) Synthesis from a set of parameter contours stored in the memory. These parameter contours, in turn, were obtained from the representation of the basic units after being modified by the rules. This requires memory for buffers, but the rule activation process is flexible since it is not tied with the synthesis process. The approach we followed is a combination of both. Basically, the excitation parameters (pitch and gain) are stored in buffers (for a full sentence or clause) and are modified by the activation of appropriate rules. The system parameters (LPC and formants) are retrieved for each basic unit and synthesized after application of appropriate rules. The block diagram is shown in Fig. 4.29.

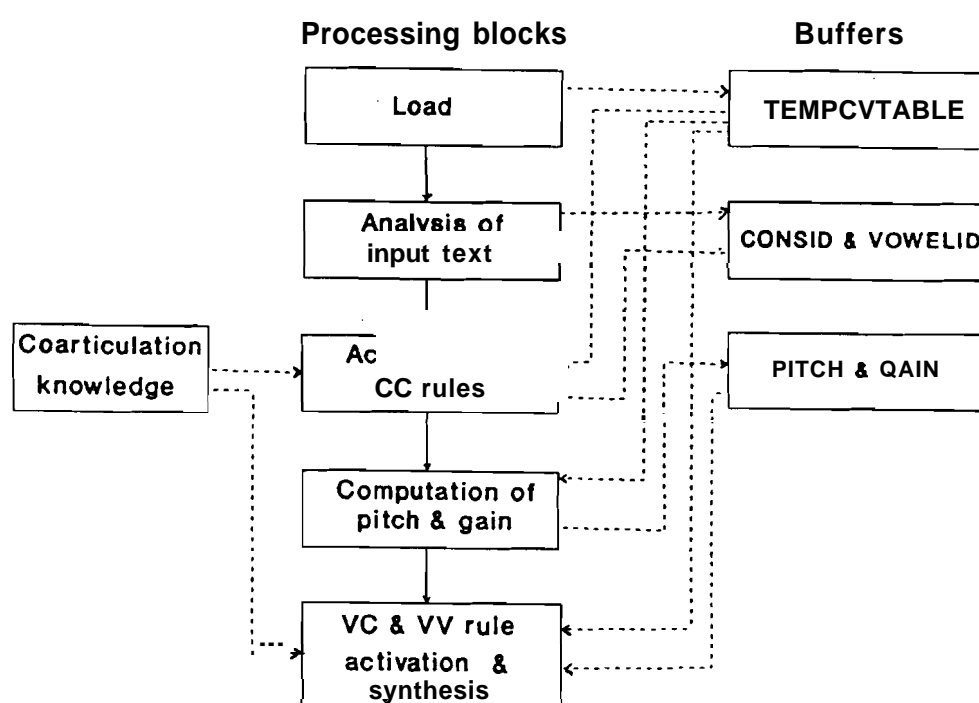


Fig. 4.29 Knowledge activation and synthesis scheme

The text is processed sentence by sentence. The prestored representations of **all** basic units in the sentence are loaded into a buffer **TEMPCVTABLE** so that the **parameter** modifications by the subsequent processing blocks can be done on these. Analysis of the input text is done to decide the phonetic context of the basic units in the text. This information is stored in the buffers **CONSID** and **VOWELID** and is used by the rule activation process. The cluster consonant rule activation process scans the text analysis information and whenever the context of some rule is matched, the rule is activated to modify the default parameters of the concerned basic unit in **TEMPCVTABLE**. Next the computation of pitch and gain contours are done using the information in the **TEMPCVTABLE**. These are stored in buffers **PITCH** and **GAIN** so that the gain and intonation rules can be activated on them before the synthesis process. The final step is the synthesis along with the activation of transition rules. The **VC**, **W** and nasalization rules are activated in this step.

4.4.2 Representation and activation of coarticulation knowledge

The straight forward way to represent these rules is by using the productions (IF-THEN rules). The activation can, then, be done using an inference mechanism, which searches for rules which satisfy the context decided by the text and then fires all of them in appropriate order. The above scheme involves search for rules for a given 'context. The following observations help us to use slightly modified rule representation and activation scheme which avoids the search which is found to be **unnecessary**. The first is that the context in which a rule is to be applied is decided completely by the basic units involved in the transition. This means that we need no complex state space search but a simple scanner which identifies the junctions between basic units one by one sequentially. The second fact is that there is a correspondence between junctions (between basic units) and rules, i.e., the junction decides the rule to be applied which is unique to that junction. This allows direct access of the rules.

4.4.2.1 Analysis of input text

Analysis of the input text is done to decide the phonetic nature of the basic units in the text. This information forms the context for the activation of the coarticulation rules. The data structures used to store this information are two arrays **CONSID** and **VOWELID**. First one stores the consonant identification code and the second the vowel identification code of each of the basic units in the input text. The assignment of the

consonant identification code is based on the consonant features, namely place and manner of articulation (refer to the consonant classification in Table.4.2). The vowel quality decides the vowel identification code. These are given in the Tables 4.15 and 4.16.

Table 4.15 The assignment of identification codes for consonant classes

Consonants	Consid
k, kh, g, gh	1
c, ch, j, jh	2
ʈ, ʈh, ɖ, ɖh	3
t, th, d, dh	4
p, ph, b, bh	5
s	6
ʃ	7
r	8
l	9
ŋ	10
n	11
m	12
ɣ	13
v	14
not used	15

Table 4.16 The assignment of identification codes for vowels

Vowel	Vowel id
a	1
e:	2
i	3
o:	4
u	5

4.4.2.2 Representation and activation of cluster consonant rules

These rules modify only the duration and gain of the **basic unit**. **This** modification is done in the basic unit representations stored in the buffer TEMPCVTABLE before

it is used for synthesis. Since the number of rules is very few (only three), they were **inbuilt** into a procedure which scans the text analysis information and activates the rules. When it is required to add more rules of this type, we can replace this procedure by an inference engine and a set of explicit rules without affecting other modules.

4.423 Representation and activation of rules which modify formants

The VC and VV rules and the nasalization rules come under this group. These rules are activated along with the synthesis process. The data structure used to represent the VC **coarticulation** rules is a table **VCTABLE**. Each slot in the **VCTABLE** is used to represent a distinct VC rule. This table is two dimensional and hence is accessed using two indices, the vowel index (vowelid) and the consonant index (consid). The range of **vowelid** is [1-5] and that of **consid** is [I-151]. Therefore the size of the table is 75. Each VC rule is stored in the table slot whose one index is the consonant identification **code** of the C in the VC and the other index is the vowel identification code. This implies that given a VC junction, we can directly access the VC rule to be applied. In this way the search in the rule activation process is avoided. The format in which a VC rule is stored is given below.

per_F1	-	percent change in the first formant from its steady value
per_F2	-	“ second ”
per_F3	-	“ third ”
per_F4	-	“ fourth ”
TRANSDUR	-	duration of formant transition in frames of 6.4 msec.

per_F4 is not used currently since we use only first three formants for synthesis.

The rule activation and synthesis module (see figure 4.29) synthesizes the basic units in **TEMPCVTABLE** one at a time. The **TEMPCVTABLE** is scanned from beginning to end and for each basic unit the parameters are retrieved from the **TEMPCVTABLE** entry. Next, using the phonetic information of both the current and

the next basic unit (these had been made available already in buffers **CONSID** and **VOWELID** by the text analysis phase) the type of junction between the current and next units is determined. If it is a VC junction, the corresponding VC rule is retrieved from the **VCTABLE**. If the junction type is VV, the vowel to vowel formant interpolation rule is used. The synthesis of the basic unit is then done as follows: The **LPCs** representing the initial region (if present) of the **unit** is synthesized from the prestored **LPC** parameters. The formant representing region is synthesized from the formant contours computed using the initial and steady formant values and the CV transition duration (from the **TEMPCVTABLE** entry) and the final formant values obtained using the VC or VV rule and the corresponding transition duration.

The nasalization rules are activated to nasalize the vowel in CV and VC contexts where C is a nasal. The nasalization effect is obtained by using a pole-zero pair below the first formant in the synthesis of the vowel. The nasal pole is fixed at 250 Hz throughout the vowel while the nasal zero is midway between the nasal pole and the first formant.

4.5 SUMMARY

This chapter discussed various issues in the acquisition, formulation and incorporation of the **coarticulation** knowledge into a text-to-speech system for Hindi. The coarticulation effects are studied in terms of parameters like formant frequency shifts, **durational** changes etc. The classification of various coarticulatory transition patterns between speech units were done based on the articulatory similarity of the transitions. Based on this a fixed number of contextual rules were formulated which specify the transition patterns for various contexts. Various issues in the incorporation of the coarticulation knowledge into the text-to-speech system were addressed. A knowledge representation scheme using tables was **proposed**. The knowledge activation were done along with the synthesis process.

Chapter 5

TESTING AND EVALUATION

5.1 INTRODUCTION

Since the database of basic units is constructed by manual editing, it has to be tested and tuning of the parameters has to be done manually to improve the quality of the units. The effectiveness of the coarticulation knowledge incorporated to the system as was discussed in Chapter 4 has to be assessed. This chapter discusses some procedures used to accomplish the above tasks. Section 5.2 discusses the testing of the basic units. Section 5.3 describes the testing and evaluation procedures used for various coarticulation rules and section 5.4 is about evaluation of quality improvement at higher levels like sentences and paragraphs.

5.2 TESTING OF THE BASIC UNITS

We have about **350** basic units in the system. Each of these was coded using various speech parameters according to the representation method described in Chapter 3. The testing of the basic units is done for the following purposes:

- (i) To assess the intelligibility of all of the basic units in **isolation**.
- (ii) To **verify** the representation of each of the units against inconsistencies.

One of the sources of inconsistency is the fact that the basic units are represented by using many control parameter values and data values. These values were derived heuristically by analyzing natural speech data. The dynamic nature of these parameters are limited by these fixed values. The testing of the basic units is done using the same utility which was used to construct its representation. This has got the following facilities to accomplish this task interactively:

- (i) Speech synthesis and playback
- (ii) Menu-driven editing of the representation
- (iii) Graphic display of the synthesized speech and the parameter contours.

Using this utility, each of the basic units in the database was synthesized and listened to carefully. Adjustments of various control frame numbers, durations, parameter values (pitch, gain, formant values etc.) are done, if necessary, and the unit is again synthesized. This is repeated until we get satisfactory **intelligibility** of the basic unit. Special care was taken to give proper CV formant transition pattern, initial silence or voicing duration, burst frame number, vowel onset frame number etc.

The testing process verified by synthesis representations of all the basic units. Regarding intelligibility of the synthesized basic units, most of them were satisfactory and comparable to that of the actual or LP synthesized units. The problematic units were the retroflex stops and the nasals. The problem with some of the retroflex sounds was in reproducing the CV transition which was more complex than a mere formant transition. The intelligibility of these units was less compared to the corresponding natural ones. In the case of nasals, the presence of zeros (antiformants) in the spectrum is very significant in deciding the quality of the unit. Though we model this using antiresonators in the formant synthesis system, the quality is not as good as the natural ones.

5.3 TESTING THE COARTICULATION RULES

The coarticulation rules incorporated to the system were tested to (i) verify their correct activation and (ii) evaluate their perceptual significance. This was done by both experimentally and perceptually.

53.1 Experimental testing

Experimental testing of the rules is done by simulating the context in which the rules would be activated and then analyzing the synthesized speech by signal processing. This was done to verify that each of the rules served its purpose when the appropriate context was present. The classes of rules tested were: (i) VC transition rules (ii) VV transition rules and (iii) CC or cluster consonant rules.

(i) Experimental testing of VC rules

The context of the VC rules are simulated by taking the sequences of basic units of two, where first is an isolated vowel and second is an isolated consonant. These are then synthesized and the synthesized speech is analysed to extract the formant transition

pattern. The formant transition pattern is then checked against the specification of the concerned rule and is verified if it is identical. The testing has been done for each of the VC rules. Fig.5.1 shows the formant contours extracted from the synthesized VC sequences against the schematic transitions generated from the corresponding VC rules.

(ii) Experimental testing of V-to-V transitions

We have formulated one general rule for bringing about vowel to vowel transitions across VV junctions in words. This rule covers all vowel to vowel transitions in Hindi. Hence it is necessary to test this rule for various VV junction allowed in Hindi. Various vowel sequences in Hindi are formed and then synthesized. The formant extraction of the synthesized speech is done and the transition patterns and their durations are verified. Fig.5.2 shows the formant contours extracted from some synthesized VV sequences.

(iii) Experimental testing of CC rules

The following are the cluster consonant rules tested:

- (i) Gemination rule
- (ii) Lengthening before a glide rule
- (iii) Word final cluster release rule

The first two of these rules modify only the durations of speech segments and the third one adds a release segment (vowel like segment) to the word final cluster consonant. Words containing these clusters are synthesized. The changes at the waveform level (i.e., changes in duration in the case of the first two rules and the addition of the release segment in the case of the third rule) are verified in each case.

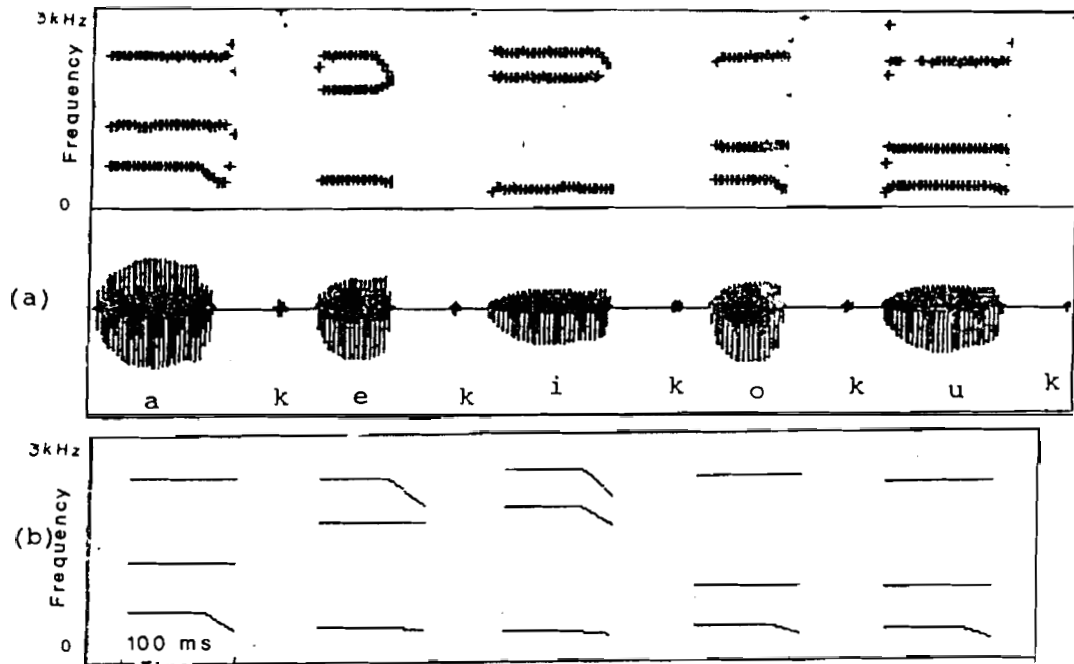
53.2 Perceptual testing of the rules

The perceptual effects of various rules are tested as explained below.

(i) Perceptual testing of VC rules

The same synthesized speech that was used for the experimental testing is used. The VC sequences synthesized with and without the transitions incorporated were listened in pairs. It is seen that even without the following consonant in the play back (i.e., only the vowel region of the VC is played to the listener), in most of the cases we

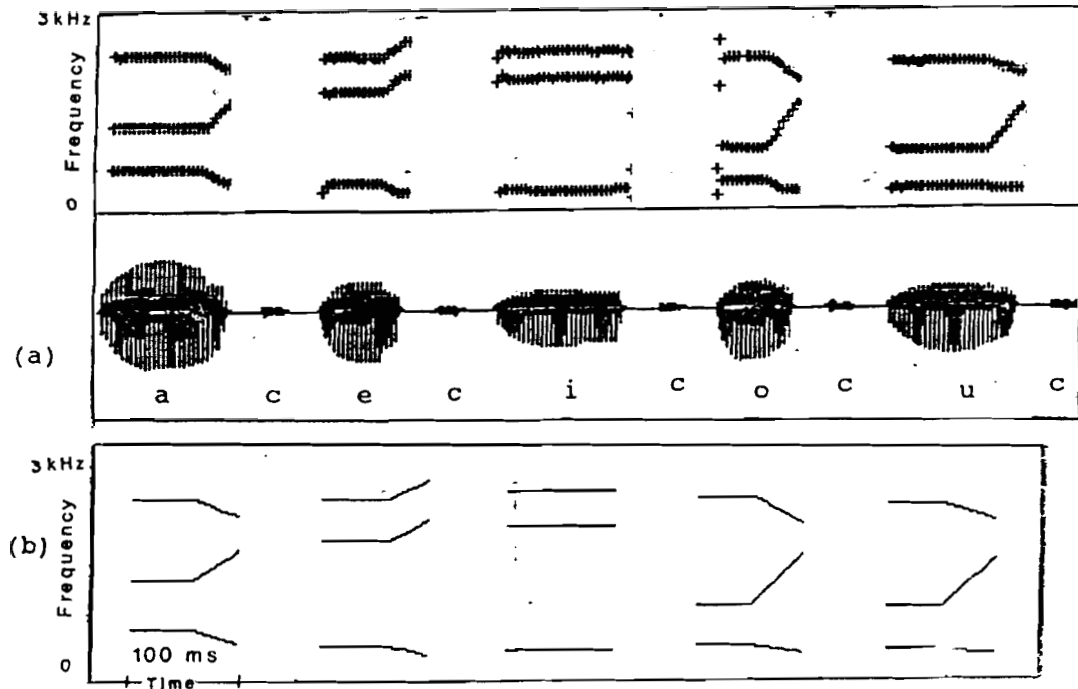
Fig. 5.1 Formant contours extracted from synthesized VC sequences



i. VC sequences involving /k/ (क)

(a) Formant contours extracted from synthesized speech

(b) Formant transition patterns specified by rules

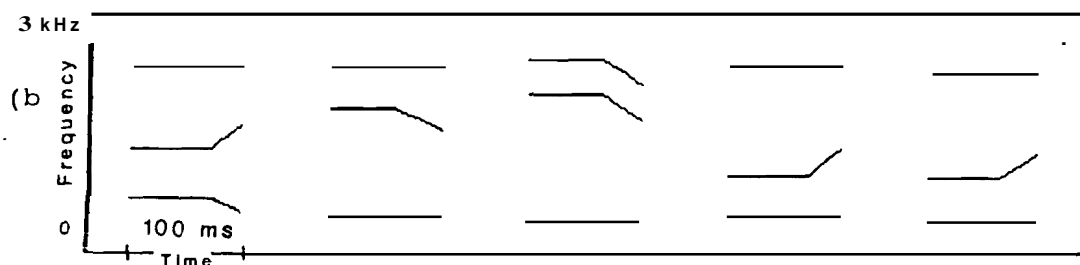
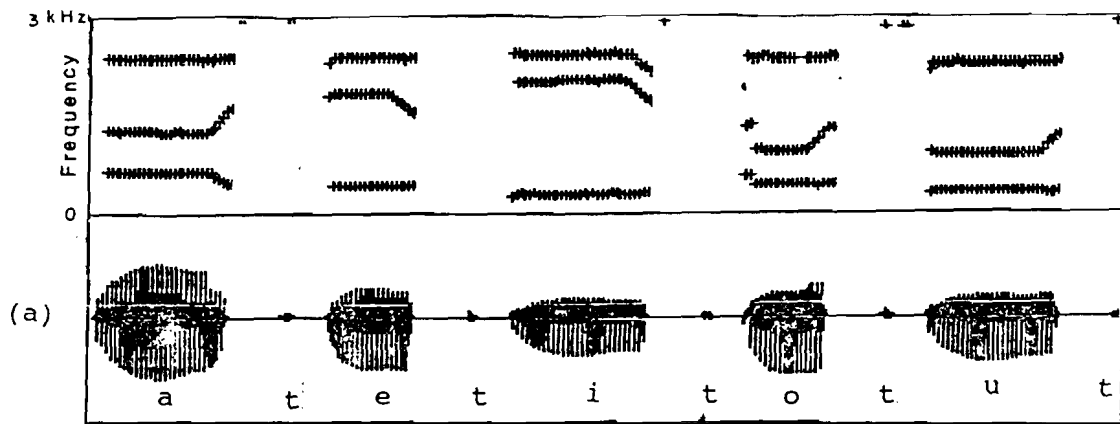


ii. VC sequences involving /c/ (च)

(a) Formant contours extracted from synthesized speech

(b) Formant transition patterns specified by rules

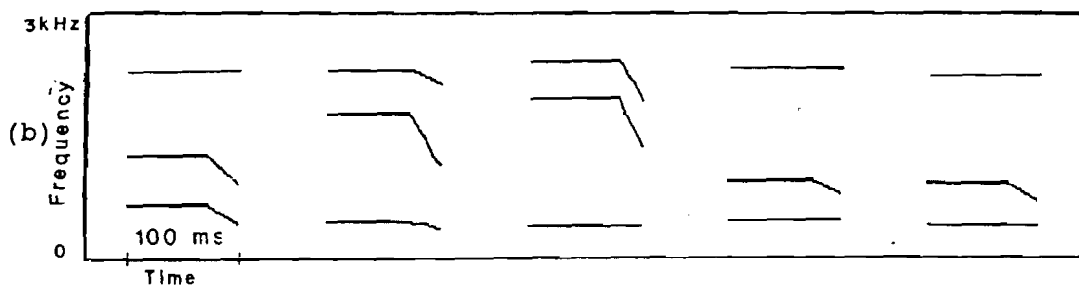
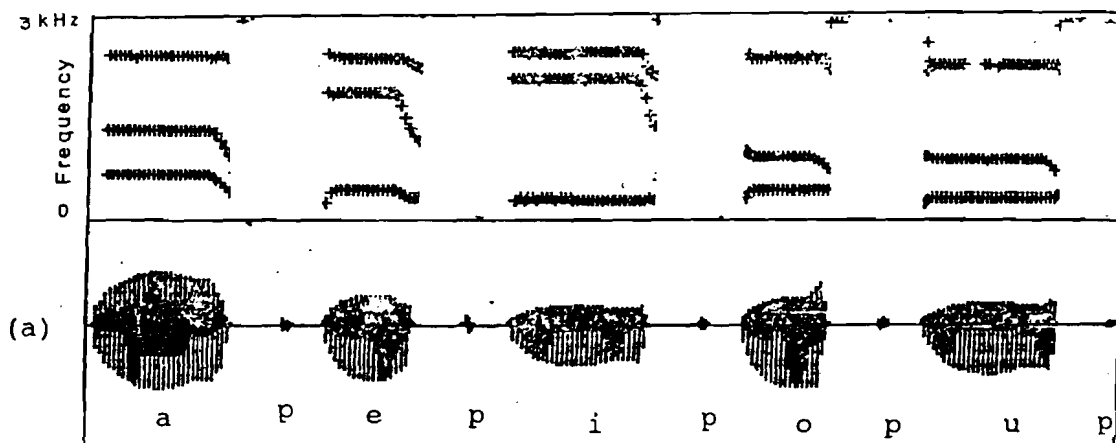
(P.T.O)



iii. VC sequences involving /t/ (π)

(a) Formant contours extracted from synthesized speech

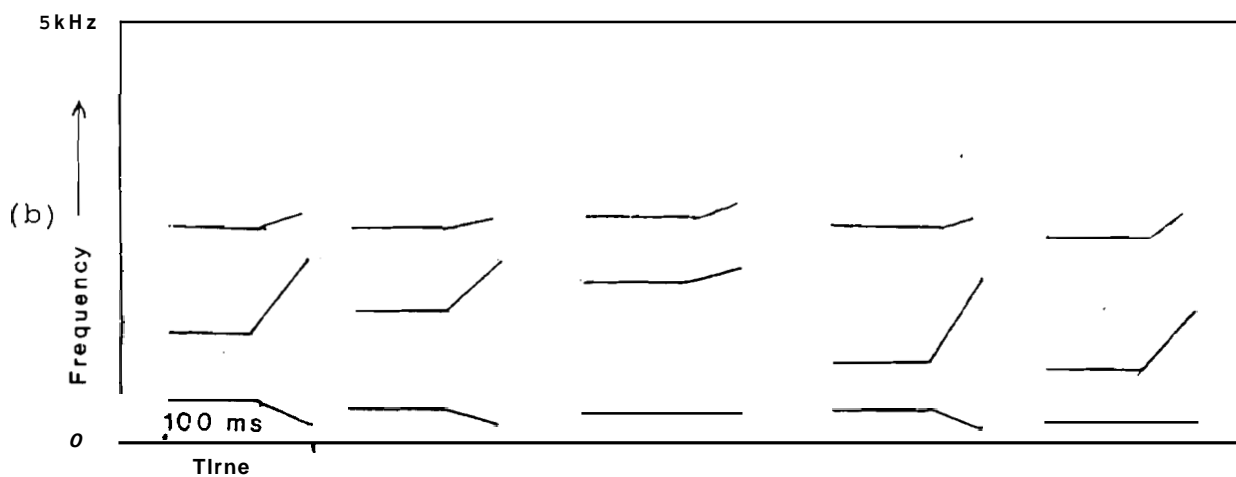
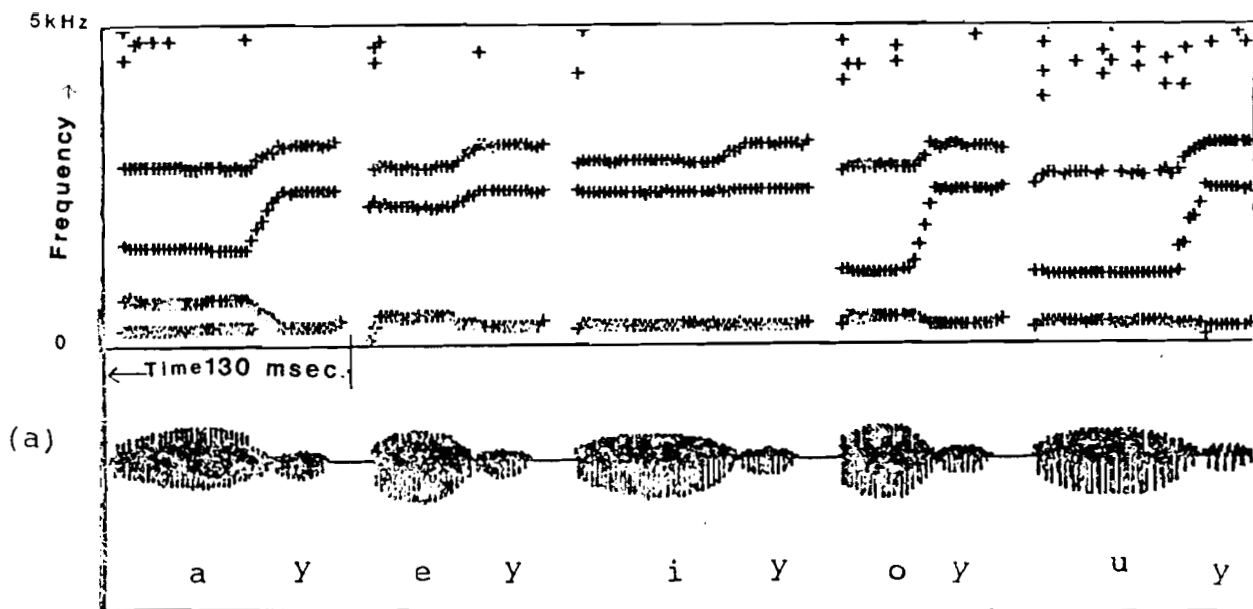
(b) Formant transition patterns specified by rules



iv. VC sequences involving /p/ (φ)

(a) Formant contours extracted from synthesized speech

(b) Formant transition patterns specified by rules (P.T.O)



v. VC sequences involving /y/ (य)

(a) Formant contours extracted from synthesized speech

(b) Formant transition patterns specified by rules

perceive the effect of the following consonant. This shows the significance of transitions in speech perception.

(ii) Perceptual testing of W and CC rules

Words containing various VV and CC junctions were synthesized both with and without the application of these rules. The improvement in perceptual quality due to rules was verified.

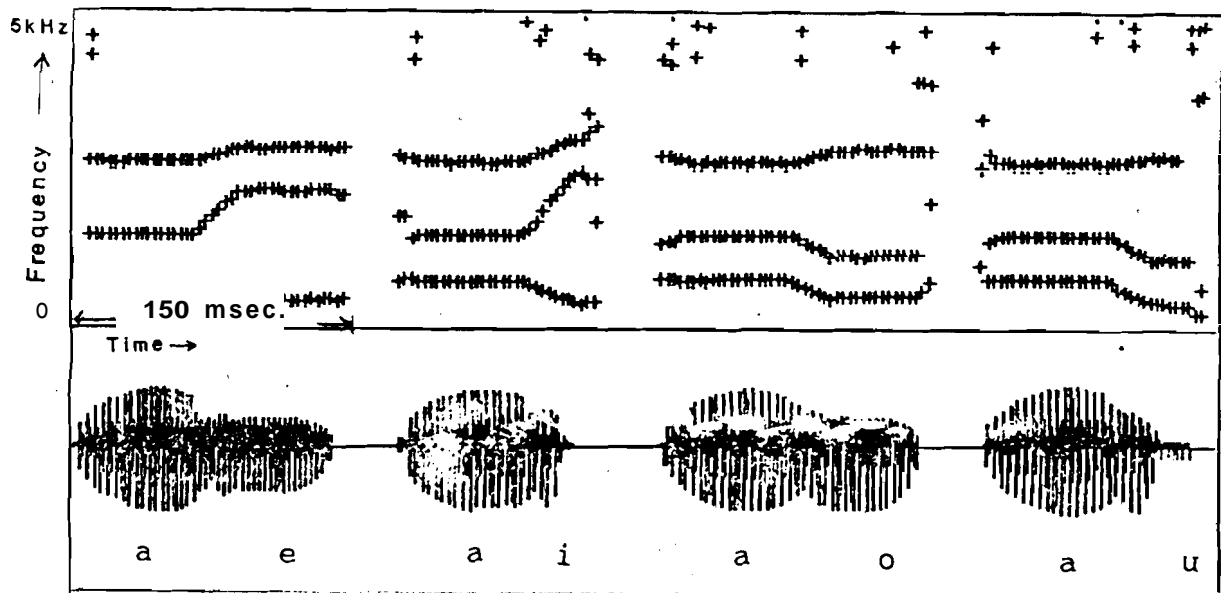


Fig. 5.2 Formant contours extracted from synthesized VV sequences

5.4 EVALUATION AT HIGHER LEVELS

In the preceding sections we discussed the testing and evaluation of the basic units and the rules used to concatenate them. To get an overall assessment of the effectiveness of the system, the evaluation was done for higher level units such as sentences. In the following we discuss the sentence and paragraph level evaluation.

Some sentences containing vowel to vowel transitions and cluster consonants besides the usual VC transitions were **selected**. These were synthesized first without the application of **any** rules and then with the rules. In both cases, the same intonation (pitch contour) was used which was computed using an intonation model for Hindi (Madhukumar et al, 1992). The sentences were played back in pairs (one without the rules and the other with rules). In all cases the quality improvement due to the rules was readily perceived.

Some paragraphs were synthesized first without coarticulation rules and then with rules. As in the previous case, same intonation was used in both cases. These were then played and listened. It was seen that the paragraph **synthesized** with the **coarticulation** rules was more natural than the one without the rules.

5.5 CONCLUSION

The basic units were tested against possible inconsistencies in the representation. The experimental tests helped in verifying the correct activation of rules in various contexts. The perceptual tests signified the importance of individual **coarticulation** rules as well as the collective effect of the rules on higher level units such as sentences and paragraphs.

Chapter 6

SUMMARY AND CONCLUSIONS

This thesis addressed some issues in designing and implementing a text-to-speech system for Hindi, an Indian language, with an emphasis on the problem of acquisition and incorporation of coarticulation. Coarticulation can be defined **as** the influence on speech segments by the neighbouring segments due to the varying degrees of physiological constraints of the articulators in continuous speech. The coarticulation gives rise to various transitional patterns between the speech segments. The transitional aspects of speech are as important as the stationary parts of the signal for proper perception of speech.

In order to incorporate the coarticulation knowledge into a text-to-speech system, we need a synthesis model which is flexible enough to make necessary modification of speech parameters. The design issues in developing such a system for an Indian language are the choice of synthesis model, the choice of basic units and their representation. The concatenation model was chosen due to its simplicity. The characters of Hindi were chosen **as** the basic units. The various phases in the text-to-speech conversion process are: preprocessing, **parṣing**, concatenation and synthesis. Preprocessing of the input text in Indian Standard Code for Information Interchange (**ISCI**) was done to expand abbreviations and numerals to their text forms. This **was** done by using lookup tables. The parser parses the expanded text into a sequence of basic unit names. The parsing process is simple and straight forward owing to the phonetic nature of Indian languages. The parameters of the basic units are concatenated and after the modification by knowledge activation, they are synthesized to give the speech output.

The representation of basic units was done using the source-system model of speech production. The consonant region of basic units was represented using Linear Prediction coefficients and the vowel region using formants and their bandwidths. The parameters pitch and gain were used to generate the excitation signal for synthesis. The issues in constructing the basic units in this representation are the speech data

collection, analysis for parameter extraction and coding of the unit using the parameters. The speech data was collected from a native Hindi speaker. The LPC, formants, pitch and gain parameters were extracted and the fields of the representation were filled in manually for each basic unit using the above parameters.

There are two major types of coarticulation based on the direction of the coarticulation effects, namely, anticipatory coarticulation or forward coarticulation and carry-over coarticulation. Coarticulation takes place between consonants (within clusters), between vowels and between the consonant and the following vowel and vice versa. The coarticulation phenomena were also classified in terms of the articulators responsible for it. Lingual coarticulation is concerned with the function of tongue as the active articulator. Nasalization is concerned with interaction of oral and nasal cavities in speech production. Vowels are nasalized in the context of nasal sounds and this process is rather a rule than exception. The major factors that affect coarticulation are the language specific constraints and external factors like speaking rate and style.

Classification of various coarticulatory transition patterns was done based on the articulatory similarity of the transitions. Based on this a fixed number of contextual rules were formulated which specify the transition patterns for various contexts. The vowel-to-consonant (VC) and vowel-to-vowel (VV) transition rules modify vowel formants and were specified as percentage changes from steady formant values and the duration of transition. The cluster consonant (CC) rules modify duration and gain. Nasalization rules nasalize the vowel in VC and CV contexts where C is a nasal. A knowledge representation scheme using tables was used. The knowledge activation was done along with the synthesis process.

Testing of the basic units was carried out interactively against inconsistencies in the representation. Testing of the coarticulation rules was done to verify the correctness of the rule activation. The perceptual evaluation indicated the importance of individual coarticulation rules as well as the collective effect of the rules on higher level units such as sentences and paragraphs.

Further improvements in synthesized speech quality can be made by (i) using an improved consonantal synthesis method, (ii) incorporating the coarticulation effects in cluster consonants at the spectral level and (iii) incorporating stress rules.

Appendix 1

VAXSTATION DETAILS

This appendix describes the hardware and software support in the VAXSTATION IVGPX system, on which the text-to-speech system is implemented. The VAXSTATION system provides an environment for performing signal processing work. The VAXlab system is a combination of hardware and software components that creates the environment that the LabStar software requires. The VAXlab system can be used to control the real time hardware which consists of the A/D converter, the D/A converter and a real time clock. The LabStar software actually provides a set of routines to perform real time I/O using the VAXlab hardware. The following two sections describe the VAXlab hardware and the LabStar software.

A1.1 VAXlab Hardware for I/O Support

The AAV1I-D is a two-channel 250-kHz digital-to-analog (D/A) converter with direct memory access (DMA). ADV1I-D is a 50-kHz analog-to-digital (A/D) converter with programmable gain and DMA. The KWV11-C clock module is used as a steady frequency source for the A/D and D/A devices. File I/O, a LabStar module device, moves data to a disk file using Queued Input Output (QIO). In QIO the user program queues buffers to the device for continuous processing of data. The device moves the data directly to disk using block I/O. As each file is read or written in blocks of ,512 bytes each, the transfer is very fast.

When the A/D and the D/A devices are set to do continuous Direct Memory Access (DMA), the DMA hardware runs continuously instead of stopping at the end of each buffer. The DMA can run at top speed without interruptions because it is confined to a 64K-byte block of memory that it wraps around. All the software has to do is to keep filling or emptying the buffers as fast as the DMA empties or fills them. We have used continuous DMA for the analog to digital conversion.

A1.2 LabStar Software for I/O Support

The LabStar Input Output (LIO) routines provide two types of interfaces: (a) synchronous read/write I/O and (b) asynchronous queued I/O. Synchronous I/O enables the user program to transfer a set of values to the device with one routine call. The routine call stops the program until the I/O completes. Asynchronous I/O enables the user program to queue several sets of values to be transferred. The program continues execution during I/O operations, enabling I/O operations to continue on one or more devices simultaneously. Asynchronous I/O has been used in the speech input/output using A-DID-A devices.

Each asynchronous I/O device has a device queue and a user queue. The user program puts a buffer in the device queue to send it to the device. The device processes the buffer and puts the buffer in the user queue to return it to the program. **LIO\$ENQUEUE** and **LIO\$DEQUEUE** are the routines for accomplishing this. With devices set for asynchronous I/O, a program can set a device to forward completed buffers to another device. When the first device completes a buffer it immediately enqueues the buffer to the second device.

Appendix 2

AN EDITOR FOR BASIC UNIT REPRESENTATION AND VC RULES

The interactive editor for the basic unit representation was developed to meet the following requirements:

- (i) To create the representation of basic units manually for each of the basic units
- (ii) To test the basic unit representation by synthesizing it
- (iii) To edit and test the VC formant transition rules

The basic unit representation, as seen in chapter 3, contains several data and control fields. This is stored as a text file in which each basic unit has got an entry for it. The text-to-speech conversion program first loads this file into a table in memory and then uses it for synthesizing the basic units. This file is updated by the editor to reflect the changes made while editing the units. The VC formant transition rules also are stored in a text file which is updated by the editor. This file also is loaded into memory by the text-to-speech program.

The main menu of the editor contains the following options:

1. Synthesize a basic unit
2. Edit a basic unit
3. **Modify a VC rule**
4. Display synthesized speech
5. Display Excitation signal

(1) Synthesize a basic unit

This synthesizes a basic unit (or a sequence of units) from the representation

(2) Edit a basic unit

This option allows the modification of various fields of an existing unit as well as the creation of a representation for a new basic unit. A sample screen is given below (user input is underlined>:

Enter the unit name: ka2

Name = ka2

1) F1 = 605 **Hz** 5) Fil = 546 **Hz** 9) Fd1 = 0 **Hz**

2) F2 = 1200 6) Fi2 = 1445 10) Fd2 = 0

3) F3 = 2285 7) Fi3 = 2363 11) Fd3 = 0

4) F4 = 3200 8) Fi4 = 3200 12) Fd4 = 0

13) CVtransdur = 6 14) Diphth_s_dur = 0 15) Diphth_t_dur = 0

16) Start_f_no = 21 17) No_of_fr = 15

18) Burst_f_no = 18 19) Onset_f_no = 21

20) Cons_gain = 1 21) Burst_gain = 5 22) Vowel_gain = 100

23) Pitch_fr_no1 = 0 24) Pitch-fr_no2 = 21

25) Pitch_ini_vo = 0 26) Pitch_vowel = 71

27) Bandwidth1 = 150 **Hz**

28) Bandwidth2 = 200

29) Bandwidth3 = 200

30) Bandwidth4 = 200

31) Quit

Enter 1..30 to modify , 31 to quit : 31

After modifying any of these values, the unit can be synthesized and the effect of modification can be observed by displaying the synthesized waveform as well as by listening to that.

(3) Modify a VC rule

This allows the interactive modification of various vowel to consonant transition rules. This also generates a graphic display of formant transition patterns from the rules. A sample screen is given below (user input is underlined):

Enter vowel id. (1..5): 1

Enter cons. id. (1..15): 2

1. percentage change in F1 = -33

2. percentage change in F2 = 30

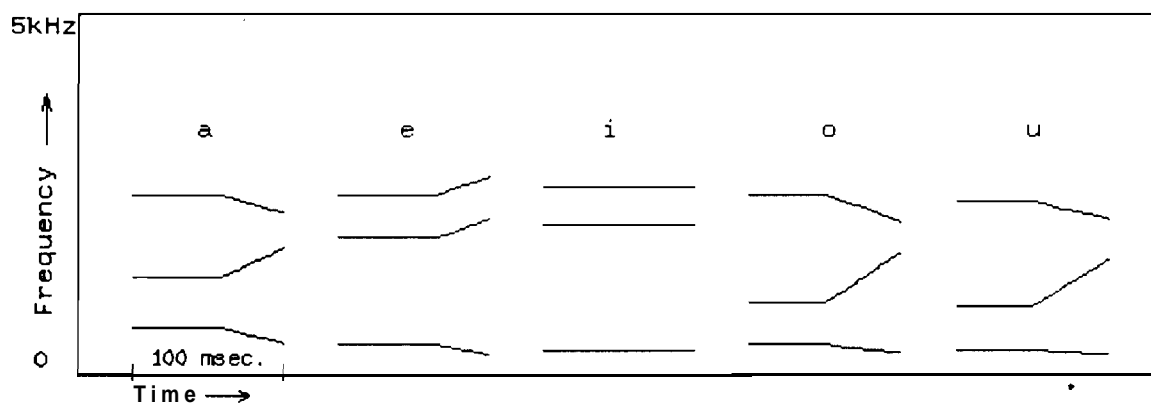
3. percentage change in F3 = -10

4. percentage change in F4 = 0

5. Trans. duration (frames) = 6

Enter 1..5 for modifying, 6 to quit: 6

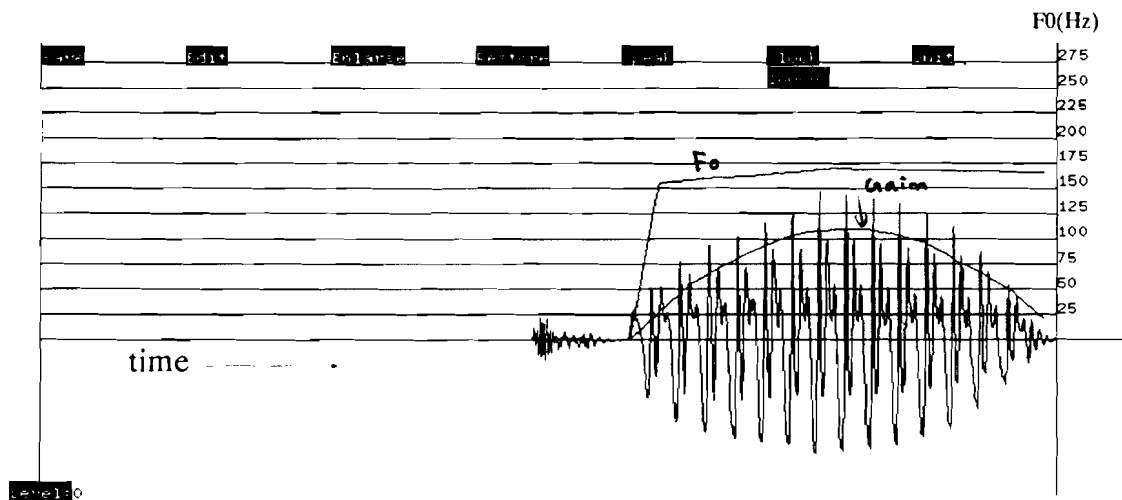
The above displays the specification of VC formant transition of the VC /ac/ (आच). The formant transition patterns from the vowels to the palatal affricates, as generated from the rules are given below.



This menu option together with the synthesis option help the tuning of the VC rules and testing them by synthesizing VC sequences.

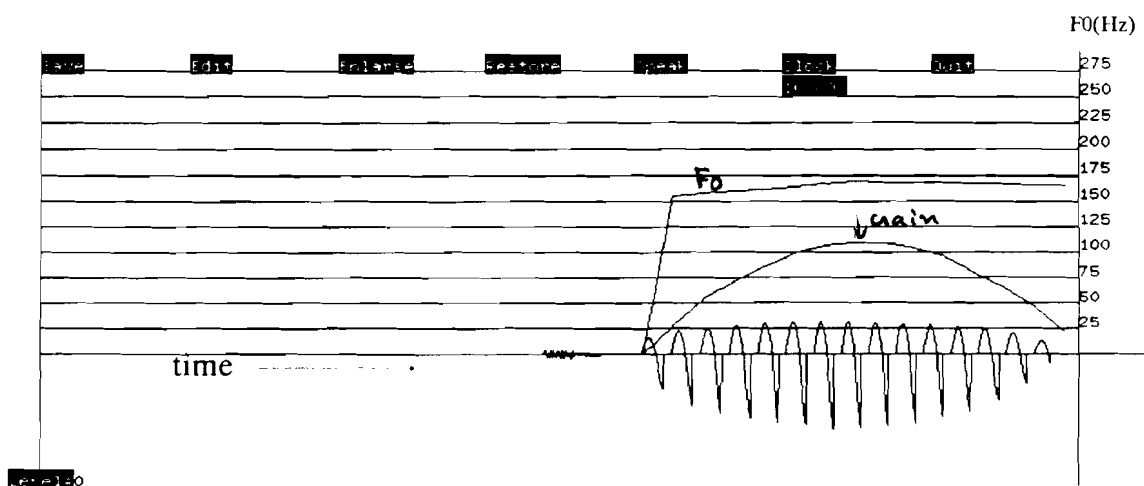
(4) Display synthesized speech

This plots the speech waveform alongwith the pitch and gain contours. A sample display is given below.



(5) Display Excitation signal

This plots the excitation signal used for synthesis. A sample display is shown below.



All these menu options together constitute an environment for editing the basic units and VC rules as well as testing them by synthesis. The collection of basic units and VC rules in the system was created using this editor.

Appendix 3

TABLE OF ISCII CODES

In this appendix we give a listing of the file TABLE.TXT. The information in this file is used by a parser. The parser is one of the modules in our text-to-speech system. The parser converts the sequence of ISCII codes into a sequence of basic units (see Section 3.3.3 for more details). The listing of the file, TABLE.TXT is as follows:

Code	Type	Index	Name	Comments

12	S	-1		CARRIAGE RETURN
32	S	-1		BLANK
33	S	-1		!
34	S	-1		“
40	S	-1		(
41	S	-1)
42	S	-1		*
44	S	-1		,
45	S	-1		-
46	S	-1		
47	S	-1		/
48	D	-1		0 }
49	D	-1		1 }

Code	Type	Index	Name	Comments
------	------	-------	------	----------

50	D	-1		ॐ
51	D	-1		3 }
52	D	-1		4 } DIGITS
53	D	-1		5 }
54	D	-1		6 }
55	D	-1		7 }
56	D	-1		8 }
57	D	-1		9 }
58	S	-1		
59	S	-1		,
63	S	-1		?
65	N	-1		ॐ CHANDRA BINDI
66	N	-1		— BINDI
68	C	0		ॐ VOWEL HEADER ; 'A'
69	C	1	K	-----
70	C	2	KH	CONSONANTS
71	C	3	G	↓
72	C	4	GH	
73	C	5	CH	
74	C	6	CHH	

Code	Type	Index	Name	Comments
------	------	-------	------	----------

75	C	7	J	
76	C	8	JH	
77	C	9	TT	
78	C	10	TTH	
79	C	11	D	
80	C	12	DD	
81	C	13	NH	
82	C	14	TH	
83	C	15	THH	
84	C	16	DH	
85	C	17	DHH	
86	C	18	N	
87	C	19	P	
88	C	20	PH	
89	C	21	B	
90	C	22	BH	
97	C	23	M	
98	C	24	Y	
99	C	25	R	
101	C	26	L	

Code	Type	Index	Name	Comments
------	------	-------	------	----------

103	C	27	V	
104	C	28	SH	
105	C	29	SHH	
106	C	30	S	
107	C	31	H	
108	V	2	AA	-----
109	V	3	I	MATHRAS
110	V	3	II	↓
111	V	5	U	
112	V	6	UU	
115	V	7	E	
116	V	8	EI	
118	V	9	O	
119	V	10	OU	-----
120	M	-1	̣ HALANTH	
121	S	-1	(FULLSTOP)	
122	M	-1	̣ NUKTHA	

Appendix 4

LOOK-UP TABLES FOR THE PREPROCESSOR

In this appendix we give listings of two text files: **ABBR.TXT** and **NUMEXP.TXT**. The information in these files is used by a preprocessor. The preprocessor is one of the modules of our text-to-speech system. The preprocessor scans the input text for abbreviations, numbers and special symbols, and converts them to their "spoken" forms (see Section 3.3.2).

The file **ABBR.TXT** contains some abbreviations and their expansions in terms of ISCII codes. The format of the data is as follows: First the abbreviation is entered, terminated by 0. Next the expansion of the abbreviation is entered, terminated by 0. This file can be modified to include more abbreviations. The listing of **ABBR.TXT** is as follows:

ABBREVIATIONS AND THEIR EXPANSIONS

Enter codes of abbr. and expansion

(each terminated by 0)

99 RU. रु.

111

46

0

99 RUPAYE रूपये

111

87

98

115

0

87 PEL.

पै.

116

46

0

87 PEISE

पैसे

116

106

115

0

106 SAN

सन्

86

120

0

106 SAN

सन्

86

120

0

68 II.

ई.

110

46

0

68

IISVII

ई स्वी

110

106

120

103

110

0

79

DAAN.

डॉ.

108

65

46

0

79

DAAKTAR

डाक्टर

108

69

120

77

99

0

87

PAN.

पं.

65

46

0

87

PANDITH

पंडित

65

79

109

82

0

69

KI.MII.

कि. मी.

109

46

97

110

0

69

KILOMITAR

किलोमीटर

109

101

118

97

110

77

99

0

106

SE.MI.

से. मी.

46		
97		
110		
46		
0		
106	SENTIIMIITAR	सेंटीमीटर
115		
86		
120		
77		
110		
97		
110		
77		
99		
0		

The file **NUMEXP.TXT** contains information about the numbers (0 to 99) and their expansions in terms of ISCII codes. Apart from these 100 numbers, the ISCII codes corresponding to a few words are also stored in this file. The information in this file is used to convert any number to its "spoken" form. The format of the data in this file is as follows: Each number and the number of ISCII codes in its spoken form are entered in one line separated by blanks. The next line contains the ISCII codes (in sequence separated by blanks), corresponding to the spoken form of the number. The listing of **NUMEXP.TXT** is as follows:

NUMBERS AND THEIR EXPANSIONS

Each record consists of the number, size of expansion,
and the ISSCII codes of the expansion (on the next line)

0	5		
	104 112 86 120 98	SHUUNYA	शून्य
1	3		
	68 115 69	EK	एक
2	2		
	84 118	DHO	दो
3	3		
	82 110 86	THEEN	तीन
4	3		
	73 108 99	CHAAR	चार
5	4		
	87 108 65 73	PAANCH	पाँच
6	2		
	74 115	CHHE	छे
7	3		
	106 108 82	SAATH	सात
8	3		
	68 108 78	AATTH	आठ
9	2		

	86 119	NOU	नौ
10	2		
	84 106	DHAS	दस
.			
.			
.			
100	2		
	106 119	SOU	सौ
101 5			
	107 75 122 108 99	HAZAAR	हजार
102 3			
	101 108 70	LAAKH	लाख
103 4			
	69 99 118 79	KAROD	करोड
104 6			
	84 104 97 120 101 103	DHASHAMLAV	दशमलव
105 4			
	87116106115	PEISE	पैसे

Appendix 5

HINDI CONSONANTS AND VOWELS

In this appendix, we list the consonants and vowels of Hindi. For each character, its name (as followed in our system) and its phonetic transcription (which is universal) are given.

Index	Character	Name	Phonetic transcription
-------	-----------	------	------------------------

Consonants:

0	अ		
1	क	K	k
2	ख	KH	kh
3	ग	G	g
4	घ	GH	gh
5	च	CH	c
6	छ	CHH	ch
7	ज	J	j
8	झ	JH	jh
9	ट	TT	t
10	ठ	TTH	ṭh
11	ड	D	d

12	ढ	DD	dh
13	ण	NH	n
14	त	TH	t
15	थ	THH	th
16	ड	DH	d
17	ध	DHH	dh
18	न	N	n
19	प	P	P
20	फ	PH	ph
21	ब	B	b
22	भ	BH	bh
23	म	M	m
24	य	Y	Y
25	र	R	r
26	ल	L	l
27	व	V	v
28	श	SH	š
29	ष	SHH	š
30	स	S	s
31	ह	H	h
32	ज	Z	z

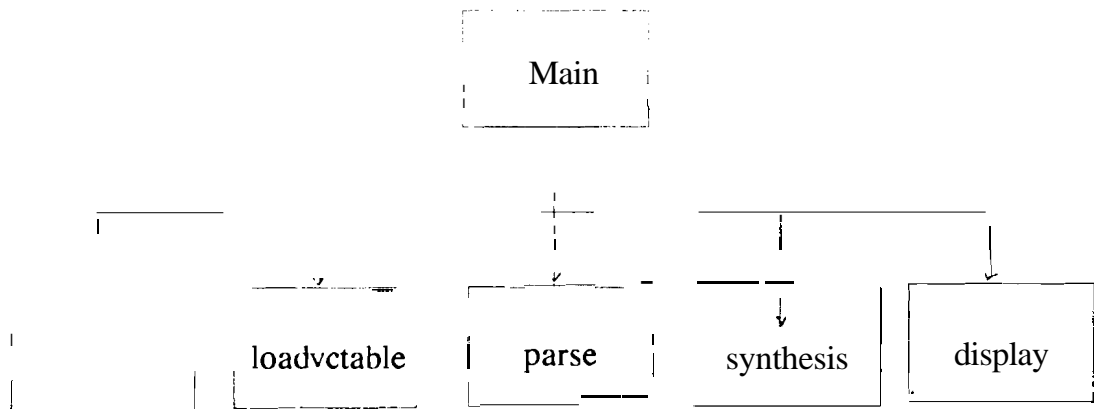
Vowels:

0	ˊ		
1		A	a
2	ˈ	AA	a:
3	ˆ	I	i
4	ˆ	II	i:
5	ˊ	U	u
6	ˊ	UU	u:
7	ˆ	E	e:
8	ˆ	EI	ai
9	ˆ	O	o:
10	ˆ	OU	au

Appendix 6

ORGANIZATION OF THE TEXT-TO-SPEECH CONVERSION SOFTWARE

The text-to-speech system software is organized in a modular way based on the function. The various modules are described with the help of an organization chart given below.



The main program has got five different modules to accomplish various tasks.

1. loadvtable: This loads all basic units from a file to a table in memory. This table is used later by the synthesis module.

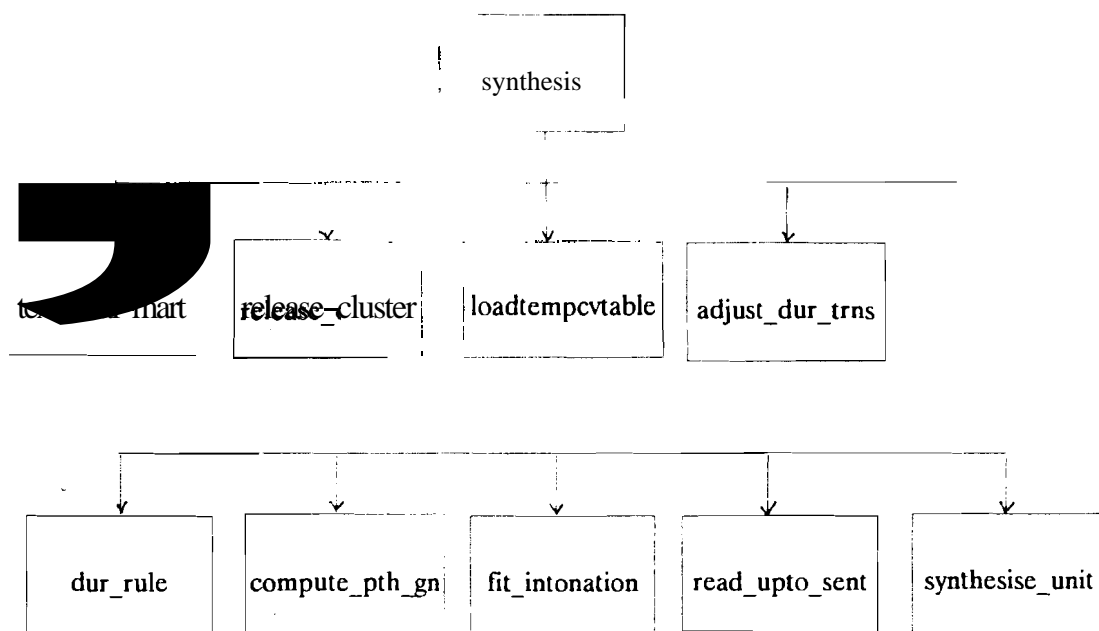
2. loadvtable: This loads all vowel to consonant rules from a file to a table in memory. This table is used later by the synthesis module.

3. parse: This module reads the input Hindi text in ISCII format, parses it into a sequence of basic units and places the output in a text file called '**fname.txt**'.

4. synthesis: This module synthesizes the sequence of basic units in the file '**fname.txt**' after applying various rules for naturalness.

5. display: This module displays the synthesized speech and pitch and gain contours. It also does the digital to analog conversion.

Of the above, the synthesis module does the rule activation and synthesis process. It has got the following organization:



1. text-anal-coart: This module performs the analysis of the input sequence of basic units to decide their consonant and vowel types. This information is stored in buffers and used later by the rule activation and synthesis module.

2. release-cluster: This module does some modifications in the sequence of basic units to enable the cluster consonant rule activation. This modified sequence of units is used by the next module.

3. loadtempcvtable: This module copies the representation of all units in the basic unit sequence into a buffer TEMPCVTABLE. This table gets modified by some of the rules and is finally used for synthesis.

4. adjust-dur-trans: This module performs the duration and gain adjustments by applying the cluster rules. This also adjusts the gain and duration across the VC and VV transitions.

5. dur_rule: This is the durational rule activation module. It modifies the inherent duration of the basic units in the TEMPCVTABLE contextually.

6. compute_pth_gain: The pitch and gain contours are computed from the TEMPCVTABLE. These, after modification by rules, are used for the generation of the excitation signal.

7. fit-intonation This module does the activation of the intonation knowledge to modify the pitch contour to get appropriate intonation for the synthesized speech.

8. read-upto-sent Loads the LP coefficients of all basic units from the corresponding files in disk into an array in memory. This array is used by the next module for synthesizing the LP represented region of basic units.

9. synthesise-unit Synthesize the units in TEMPCVTABLE one by one after applying the VC and VV transition rules.

REFERENCES

- J. Allen, "Synthesis of speech from unrestricted Text," Proc. IEEE, 4, pp.433-442, 1976.
- J. Allen, M. Hunnicutt and D.H. Klatt, From Text to Speech: The *MITalk* System, Cambridge University Press, Cambridge, 1987.
- A.P. Benguerel and H.A. Cowen, "Coarticulation of upper lip protrusion in French," *Phonetica*, vol.30, pp.41- 55, 1974.
- D.G. Childers, Ke Wu, D.M. Hicks and B. Yegnanarayana, "Voice conversion," Speech communication, vol.8, pp.147-158, 1989.
- P. Delattre, A. Liberman and F. Cooper, "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am., vol.27, pp.769-773, 1955.
- G. Docherty and L. Shockey, "Speech Synthesis", In Aspects of Speech Technology, ed. by M. Jack and J. Laver, Edinburgh University Press, pp.144-183, 1988.
- G. Fant, "Acoustic analysis and synthesis of speech with applications to swedish," *Ericsson Technics*, No.1, 1959.
- G. Fant, Acoustic theory of speech production, Mouton, The Hague, Paris, 1960.
- C.A. Ferguson, "Universal tendencies and normal nasality," In *Nasalfest: papers from a symposium* on nasals and nasalization, C.A. Ferguson, L.M. Hyman and J.J. Ohala (eds), Stanford: Language Universal Project, 1975.
- T. Gay, Articulatory movements in VCV sequences, J. Acoust. Soc. Am., vol.62, pp.183-193, 1977.
- G. Heike, R. Greisbach and B.J. Kroger, "Coarticulation rules in an articulatory model," J. Phonetics, vol.19, pp.465-471, 1991.
- W.L. Henke, Dynamic articulatory model of speech production using computer simulation, Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, Mass., 1966.

- J.N. Holmes, "Formant synthesizers: cascade or parallel ?," *Speech Communication*, vol.2, pp.251-273, 1983.
- J.N. Holmes, *Speech Synthesis and Recognition*, Van Nostrand Reinhold (UK), Cornwell, 1988.
- A.S. House and K.N. Stevens, "Analog studies of the nasalization of vowels", *J. of Speech and Hearing Disorders*, vol.21, pp.218-232, 1956.
- Y. Ishikawa and K. Nakajima, "Neural network based spectral interpolation method for speech synthesis by rule," *Eurospeech*, vol.1, pp.47-50, 1991.
- R.D. Kent and K.L. Moll, "Tongue body articulation during vowel and diphthong gestures," *Folia Phoniatrica*, vol.24, pp.286-300, 1972.
- D.H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol.67, pp.971-995, 1980.
- V.A Kozhevnikov and L.A. Chistovich, "Speech: Articulation and Perception," Washinton, D.C.: Joint Publication Research Service, No.30, pp.543, 1965.
- I. Lehiste, *Suprasegmentals*, The MIT Press, Cambridge, 1970
- A.M. Liberman, "The grammars of speech and language," *Cognitive Psychology*, vol.1, pp.301-323, 1970.
- K. Lukaszewicz and M. Karjalainen, "Microphofiemic method of speech synthesis, ICASSP, pp.1426-1429, 1987.
- P.F. Mac Neilage and J.L. Declerk, "On the motor control of coarticulation in CVC monosyllables," *J. Acoust. Soc. Am.*, vol.45, pp.1217-1233, 1969.
- A.S. Madhukumar, S. Rajendran and B. Yegnanarayana, "Intonation component in a text-to-speech system for Hindi", Paper accepted for *Computer Speech and Language*.
- J. Makhoul, "Linear Prediction: A tutorial review," *Proc. IEEE*, vol.63, pp.561-580, 1975.
- S.Y. Manuel, "The role of contrast in limiting vowel-to-vowel coarticulation in different languages," *J. Acoust. Soc. Am.*, vol.88(3), pp.1286-1298, 1990.

- K.L. Moll and R.G. Daniloff, "Investigation of the timing of velar movements during speech," *J. Acoust. Soc. Am.*, vol.**50**, pp.678-684, 1971.
- K. Nagamma Reddy, "Nasalization in Telugu: an electrokymographic study," In *Studies in Dravidian and General Linguistics* ed. by B. Lakshmi Bai and B. Ramakrishna Reddy, pp.141-166, Osmania University, Hyderabad, 1990.
- M. Ohala, *Aspects of Hindi Phonology*, Motilal Banarsidass, New Delhi, 1983.
- S.E.G. Ohman, "Coarticulation in VCV utterances: spectrographic measurements," *J. Acoust. Soc. Am.*, vol.**39**, pp.151-168, 1966.
- P.E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice Hall, Englewood Cliffs, N.J, 1987.
- G.E. Peterson and H.L. Barney, "Control method used in a study of the vowels," *J. Acoust. Soc. Am.*, vol.**24**, pp.175- 184, 1952.
- R.K. Potter and J.C. Steinberg, "Toward the specification of speech," *J. Acoust. Soc. Am.*, vol.**22**, pp.807-820, 1950.
- L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, New Jersey, 1979.
- S.R. Rajesh Kumar, "Significance of durational knowledge for a text-to-speech system in an Indian language," M.S. Thesis, Indian Institute of Technology, Madras, March 1990.
- M. G. Rizet, "A rule-based segmental synthesis module for French," *Eurospeech*, vol.**1**, pp.51-54, 1991.
- D. O'Shaughnessy, "Design of a real-time French text-to-speech system," *Speech Communication*, 3, pp.233-243, 1984.
- D. O' Shaughnessy, *Speech Communication - Human and Machine*, Addison Wesley, Reading, 1987.
- S. Smith, "Vocalization and added nasal resonance", *Folia Phoniatica*, vol.**3**, pp.165-169, 1951.
- S. Srikanth, S .R. Rajesh Kumar, R. Sundar and B. Yegnanarayana, 'A text-to-speech conversion system for Indian languages based on waveform

concatenation model," Technical Report No. 11, Project VOIS, Dept. of Computer Science and Engineering, I.I.T Madras, March 1989.

R. Sriram, S.R. Rajesh Kumar, and B. Yegnanarayana, A text-to- speech conversion system for Indian languages using parameter based approach,. *Technical* Report No. 12, Project VOIS, Dept. of Computer Science and Engineering, I.I.T Madras, May 1989.

R.H. Stetson, Motorphonetics, Amsterdam: North Holland, 1951.

W. Strange, J. Jenkins and T. Johnson, "Dynamic specification of coarticulated vowels", *J. Acoust. Soc. Am.*, vol.74, pp.695-705, 1983.

D.H. Whalen, "Coarticulation is largely planned," *J. Phonetics*, vol.18, pp.3-35, 1990.

W.A. Wickelgren, "Context sensitive coding, associative memory, and serial order in (speech) behaviour," *Psychological Review*, vol.76, pp.1-15, 1969.

B. Yegnanarayana, D.G. Childers and J.M. Naik, A Flexible Analysis Synthesis system for studies in speech processing, Internal Report, Dept. of Electrical Engineering, University of Florida, Gainesville, 1984.

B. Yegnanarayana, H.A. Murthy and V.R. Ramachandran, Speech Enhancement using Group Delay Functions, Proc. International Conference on *Spoken Language* Processing, Kobe, Japan, vol.2, pp.301-304, November 1990.

B. Yegnanarayana, H.A. Murthy and V.R. Ramachandran, Speech Processing using Modified Group Delay Functions, Proc. *ICASSP* 91, Toronto, PP.945-948, March 1991.

B. Yegnanarayana and V.R. Ramachandran, "Group delay processing of speech signals," ESCA Workshop on Comparing Speech Signal Representations, Sheffield, England, pp.411- 418, April 1992.

LIST OF FIGURES

Fig. 1.1 Illustration of the coarticulation effect between speech segments

Fig. 2.1 Parts of vocal tract system involved in speech production

Fig. 2.2 Effect of nasalization on vowel [a]

Fig. 3.1 The speech synthesis model

Fig. 3.2 Phases in text-to-speech conversion

Fig. 3.3 Various excitation signals for speech synthesis

Fig. 3.4 Basic unit representation information with respect to a waveform

Fig. 4.1 Spectrogram of VC transition /a:k/ (आक)

Fig. 4.2 VC formant transitions in the context of velar stops

Fig. 4.3 VC formant transitions in the context of palatal affricates

Fig. 4.4 Spectrogram of VC transition /a:c/ (आच)

Fig. 4.5 VC formant transitions in the context of palatal fricative

Fig. 4.6 Spectrogram of VC transition /a:ʃ/ (आश)

Fig. 4.7 VC formant transitions in the context of palatal glide

Fig. 4.8 Spectrogram of VC transition /a:y/ (आय)

Fig. 4.9 VC formant transitions in the context of retroflex stops

Fig. 4.10 Spectrogram of VC transition /a:t/ (आट)

Fig. 4.11 VC formant transitions in the context of retroflex nasal

Fig. 4.12 Spectrogram of VC transition /a:n/ (आण)

Fig. 4.13 VC formant transitions in the context of alveolar fricative

Fig. 4.14 Spectrogram of VC transition /a:s/ (आस)

Fig. 4.15 VC formant transitions in the context of the lateral

Fig. 4.16 Spectrogram of VC transition /a:l/ (आल)

Fig. 4.17 VC formant transitions in the context of the trill

Fig. 4.18 Spectrogram of VC transition /a:r/ (आर)

Fig. 4.19 VC formant transitions in the context of dental stops

Fig. 4.20 Spectrogram of VC transition /a:t/ (आत)

Fig. 4.21 VC formant transitions in the context of dental nasal

Fig. 4.22 Spectrogram of VC transition /a:n/ (आन)

Fig. 4.23 VC formant transitions in the context of labio-dental glide

Fig. 4.24 Spectrogram of VC transition /a:v/ (आव)

Fig. 4.25 VC formant transitions in the context of bilabial stops

Fig. 4.26 Spectrogram of VC transition /a:p/ (आप)

Fig. 4.27 VC formant transitions in the context of bilabial nasal

Fig. 4.28 Spectrogram of VC transition /a:m/ (आम)

Fig. 4.29 Knowledge activation and synthesis scheme

Fig. 5.1 Formant contours extracted from synthesized VC sequences

Fig. 5.2 Formant contours extracted from synthesized VV sequences

LIST OF TABLES

Table 1.1 Various categories of knowledge, their domain, parameters affected and linguistic function

Table 3.1 Basic unit names of the delimiters

Table 3.2 Basic unit representation information for /ka/ (क)

Table 4.1 Basic units and corresponding domain of coarticulation

Table 4.2 Classification of Hindi consonants using the place and manner of articulation features

Table 4.3 VC transitions in the context of velar stops

Table 4.4 VC transitions in the context of palatal affricates

Table 4.5 VC transitions in the context of palatal fricative

Table 4.6 VC transitions in the context of retroflex stops

Table 4.7 VC transitions in the context of retroflex nasal

Table 4.8 VC transitions in the context of alveolar fricative

Table 4.9 VC transitions in the context of the lateral

Table 4.10 VC transitions in the context of the trill

Table 4.11 VC transitions in the context of dental stops

Table 4.12 VC transitions in the context of dental nasal

Table 4.13 VC transitions in the context of bilabial stops

Table 4.14 VC transitions in the context of bilabial nasal

Table 4.15 Assignment of identification codes for consonant classes

Table 4.16 Assignment of identification codes for vowels

LIST OF PUBLICATIONS

List of related publications

1. V.R. Ramachandran and B. Yegnanarayana, "Coarticulation rules for a text-to-speech system for Hindi", Proc. of Workshop on Speech Technology, Indian Institute of Technology, Madras, pp. 211-219, Dec. 1992.
2. B. Yegnanarayana, S. Rajendran, S.R. Rajesh Kumar, V.R. Ramachandran and A.S. Madhukumar, "Knowledge sources for a text-to-speech system in Hindi", Second Regional Workshop on Computer Processing of Asian Languages, Indian Institute of Technology, Kanpur, pp. 233-242, March 1992.
3. B. Yegnanarayana, Hema A. Murthy, R. Sundar, V.R. Ramachandran, A.S. Madhukumar, N. Alwar and S. Rajendran, "Development of a text-to-speech system for Indian languages", Frontiers in Knowledge-based Computing, Proc. of KBCS'90, Pune, pp. 467-476, Dec. 1990.
4. S.R. Rajesh Kumar, V.R. Ramachandran, A.S. Madhukumar, Hema A. Murthy and B. Yegnanarayana, "A text-to-speech system for Indian languages", Paper presented at the Seminar on Common Phonetic Matrix for Indian Languages, CIIL, Mysore, March 1990.

List of other publications

1. B. Yegnanarayana and V.R. Ramachandran, "Group delay processing of speech signals", ESCA Workshop on Comparing Speech Signal Representations, Sheffield, England, pp. 411-418, April 1992.
 2. B. Yegnanarayana, Hema A. Murthy and V.R. Ramachandran, "Speech processing using modified group delay functions", Proc. of ICASSP-91, Toronto, Canada, pp. 945-948, May 1991.
 3. B. Yegnanarayana, Hema A. Murthy and V.R. Ramachandran, "Speech enhancement using group delay functions", Proc. of ICSLP90, Kobe, Japan, pp. 301-304, Nov 1990.
-