INTONATION KNOWLEDGE FOR SPEECH SYSTEMS FOR AN INDIAN LANGUAGE

A thesis
submitted for the award of the degree of

Doctor of Philosophy
in
Computer Science and Engineering

by

Madhukumar. A. S.

Department of Computer Science and Engineering
Indian Institute of Technology
Madras 600 036, INDIA

January 1993

CERTIFICATE

This is to certify that the thesis entitled **Intonation knowledge for speech systems for an Indian language** is a bona fide work of Mr. A. S. Madhukumar, 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 Doctor of Philosophy in Computer Science and Engineering. The content of this thesis have not been submitted to any other institute or university for the award of any degree/diploma.

January 1993
IIT Madras

(B. Yegnanarayana)

ACKNOWLEDGEMENTS

I express my deep gratitude to Prof. B. Yegnanarayana, whose inspiring guidance has been a constant source of encouragement for me. 1 am grateful to him for providing a challenging research environment and stimulating research goals. His influence on the present work has been tremendous.

Many of the research issues discussed in this thesis have evolved from the discussions of speech group at IIT Madras. Rajendran, Rajesh Kumar and Ramachandran have contributed significantly to these discussions. I thank them all for their help and cooperation, without which this thesis would not have taken shape. The speech data have been collected with the help of Bharadwaj, Manish Sinha and many other graduate students of the institute, who read and provided the utterances for this thesis. I sincerely thank all of them.

All my efforts in this thesis would not have been possible without the constant encouragement and support of the faculty and staff of Department of Computer Science and Engineering, IIT Madras, especially Prof. Kamala Krithivasan and Prof. R. Nagarajan. I gratefully acknowledge their whole hearted support.

I should thank the excellent research environment in the Speech and Vision Laboratory, Department of Computer Science and Engineering, IIT Madras. I especially acknowledge many useful discussions I had with the following members of this group: Sundar, Hema, Ramana Rao, Chandra Sekhar, Chouhan, Ramaseshan, Raghu, Ravichandran, Arul Valan, Antony Ravi Singh and Udaya Kumar.

My friends and parents should not go unmentioned, who in one way or other went with me during the course of this research. I thank them for providing the moral support.

CONTENTS

ABSTRACT

To produce speech from a given text, human beings effortlessly use several knowledge sources such as prosodics, phonetics, phonology, morphology, syntax, semantics and pragmatics. Mere concatenation of the signals corresponding to the basic units of speech does not produce intelligible and natural sounding speech. The rules governing at least the prosodic knowledge (intonation, duration and intensity) are essential. Even though the prosodic knowledge by itself does not contribute directly to speech information, it helps to organize the knowledge at segmental and suprasegmental levels by incorporating proper values of pitch, gain and duration of sequences of the basic units. In this research we study the intonation knowledge, which describes variations of the fundamental frequency (F_0) with time for various speech systems such as text-to-speech, speech-to-text and speaker recognition systems for Hindi, an Indian language.

Intonation knowledge is not explicitly taught/learnt when we learn to speak. Hence it is difficult to state the rules governing the intonation for an utterance. However, intonation patterns in Hindi show some regular features. Like many other languages, F_0 contour of declarative sentences in Hindi decline gradually with time and for interrogative sentences F_0 contour rises towards the end. This backdrop declination or rising is accompanied by local falls and rises. F_0 contour gets modified across major syntactic boundaries which is called resetting of F_0 contour. The resetting is used as a marker for phrase boundaries and it is accompanied by a pause. The intonation pattern of an utterance is also affected by the phonetic factors of constituent units. Intonation knowledge for continuous speech accounts for all these factors.

Incorporation of proper intonation knowledge will enhance the intelligibility and naturalness of synthetic speech significantly. In a text-to-speech

system intonation refers to the periodicity of the source for voiced speech sounds. By exploiting the properties of F_0 contour, we are able to hypothesize some word boundaries and function words from continuous speech in Hindi. The properties of intonation patterns will also help us to differentiate one speaker from another. One advantage of using intonation knowledge for word boundary hypothesization and speaker recognition is due to robustness of intonation features even in noisy speech.

The major contributions of this thesis are as follows: (1) A model is proposed for intonation patterns in Hindi based on the type and structure of sentences, the nature of words, and the inherent phonetic properties of vowels and the contextual variations. (2) A method is proposed to incorporate intonation knowledge in a text-to-speech system for Hindi to study the significance of the knowledge in the system. (3) An algorithm is proposed to hypothesize some word boundaries for continuous speech in Hindi. The algorithm also hypothesizes a few ***function words*** (words which have only grammatical meaning) in continuous speech. (4) A speaker recognition system for small (about 10 speakers) population is developed using some robust features of intonation of a test utterance.

Chapter 1
INTRODUCTION

During production of speech, we use several knowledge sources related to intonation, stress and rhythm to convey an idea more effectively. Human beings acquired these knowledge sources without explicit training or learning. Moreover, these knowledge sources by themselves do not make up speech, but without which it is very difficult to produce intelligible and meaningful speech. Hence these knowledge sources can be viewed as *metalevel* knowledge. Acquisition and incorporation of this knowledge increase the performance of speech systems significantly. The three speech systems we consider are: (1) a text-to-speech system; (2) a speech-to-text system and (3) a speaker recognition system. Through this research, we study the significance of intonation knowledge for the speech systems in Hindi, an Indian language. The aim of this research is (1) to explore the properties of intonation patterns of continuous speech in Hindi; (2) to incorporate intonation knowledge in a text-to-speech system for Hindi; (3) to hypothesize word boundaries from continuous speech in Hindi based on the properties of intonation patterns and (4) to design and develop a speaker recognition system based on intonation knowledge.

Speech is the primary method for communication between human beings. Of the many varieties of life sharing in our world, only human beings have developed the vocal means of coding to convey information beyond a rudimentary level. Through the development of a system for speech communication between man and machine, we constitute a whole new range of communication services to extend man's capabilities, serve his social needs, and

increase his productivity. As computers become increasingly popular in nearly all segments of society, it is quite natural to consider a natural mode as medium of communication. Speech is obviously the most useful medium of communication between computers and its human users. The other possible method for representing natural communication is in text mode which can be considered as a string of conventional symbols (Allen, 1985). Text is often considered as more durable medium of communication and preserved more reliably. Hence it is widely used for both input and output of computers. But text requires specialized equipments as well as typing and reading skills which many potential users may not posses. On the other hand, speech is the communication medium most widely used between humans and therefore seems extremely natural and requires no special training. It also gives a humanizing quality to computer based systems which is highly desired. Due to these advantages, there is a growing trend towards the development of speech systems over the past three decades.

Most of the problems that computers currently solve use programs where steps of solution are defined explicitly. The conventional programs are rigid structurally, their actions are predictable in advance and they cannot handle problems that their programmers did not foresee. But as human beings, we are able to handle and frequently solve problems for which algorithms do not exist and which are characterized by ill structure, ambiguity, incomplete problem understanding, uncertainty and formidable complexity. The ability of human beings to solve such problems is almost taken for granted and is not fully understood. We are using various tools for solving problems other than unique human ability of common sense reasoning such as logic, heuristic search and the extensive use of domain knowledge. To perform natural tasks by computer, one has to program computers to exhibit similar problem solving capability, or perhaps in some cases to surpass human beings. Acquisition and incorporation of domain knowledge which include formal and empirical components, play a key role in this experiment and hence this approach can be called as a knowledge-based approach.

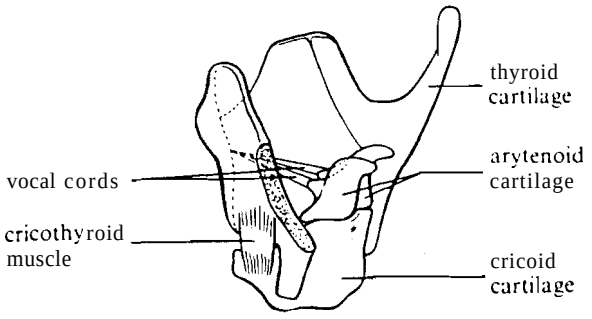
Vision and speech are two primary senses of human beings. Man learns

about his environment largely through his eyes and communication is done mainly through the voice. Both human visual system and speech production mechanism have their limitations and peculiarities. In order to incorporate features of these primary senses into the machines, we have to formulate a set of rules which considers the possibilities and limitations associated with the task, converts them into a sequence of representable form and incorporates them into a machine in some systematic fashion. Acquisition of intonation knowledge from continuous speech and incorporation of the knowledge in various speech systems demonstrate some aspects of knowledge-based systems related to man-machine communication by voice.

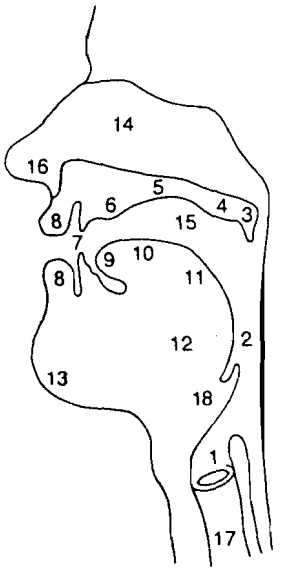
1.1 Importance of prosodic knowledge in continuous speech

Consider the production of speech. It can be viewed as a closely coordinated movements of several groups of anatomical structures. They are summarized as follows: (1) Structures which enclose the air passage below larynx (trachea): Due to the control of muscles and forces generated by the elastic recoil of the lungs, pressure is built-up below the larynx. This provides the energy for the generation of sound. (2) Structures above trachea (larynx): Position of vocal folds in larynx and their tension can be adjusted in many ways so that air can flow through the glottis with or without setting the vocal folds into vibration. If there is vibration, airflow through the glottis is interrupted quasi-periodically and thus creating an effect of modulation. (3) Structures consist of tongue, jaw, lips, velum and other components that form vocal and nasal cavities: By changing the vocal tract, one can shape the detailed characteristics of the speech sound being produced (Zue & Schartz, 1980). Fig.1.1 shows the human anatomical structures related to speech production. Fig.1.1a shows a perspective view of the larynx. Fig.1.1b shows the positions of various articulators in the vocal tract.

In this generation process the speaker does not merely articulate the successive speech sounds that make up an utterance, but simultaneously controls the vocal features such as loudness, rhythm, pitch, voice quality, etc.. These



(a)



(b)

Fig.1.1 Human anatomical structures related to speech production

(a) **A** perspective view of the larynx (’t Hart, Collier & Cohen, 1990). Larynx consists of several cartilages, the most important of which are the thyroid, the crocoid and two arytenoid cartilages.

(b) Positions of speech articulators in the vocal tract (0'Shaughnessy, 1987): (1) vocal folds, (2) pharynx, **(3)** velum, (4) soft palate, (5) hard palate, (6) alveolar ridge, (7) teeth, (8) lips, (9) tongue lip, (10) blade, (11) dorsum, (12) root, (13) mandible (jaw), (14) nasal cavity, (15) oral cavity, (16) nostrils, (17) trachea, (18) epioglottis.

features do not shape the phonetic identity of speech segments, but constitute a suprasegmental layer in the sound pattern ('t Hart, Cohen & Collier, 1990), which can be defined more globally and depend on more than the sound currently being produced. They are acoustically manifested as the variation of fundamental frequency (F_0, which is defined as the rate of vibration of vocal folds), durations of sound units and pauses, and amplitudes of the speech sounds. A typical speech waveform is shown in Fig.l.2. The word consists of the nasal */m/* followed by the vowel */a:/*, the unvoiced fricative */s/*, and finally the unvoiced stop */k/*. The figure illustrates typically the relative amplitudes of voiced and unvoiced sounds and the periodicities in voiced speech. The periodicity exhibited in voiced speech waveform is described in terms of pitch, expressed by the pitch period To or pitch frequency Fo.

Suprasegmental features are perceptual aspects of the speech signal and are

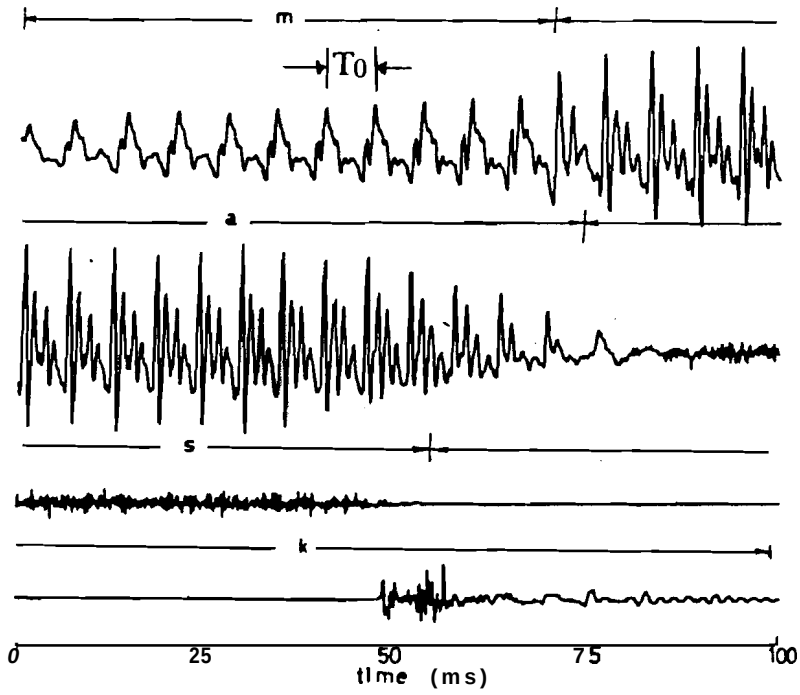


Fig.l.2 Speech waveform for the word */ma:sk/*

independent of segmental features. They are generally referred to intonation, stress and rhythm of the speech signal. They are also called as ***prosodic knowledge.*** Prosodic knowledge enhances the speech signal by carrying different higher level knowledge sources related to phonology, morphology, syntax, semantics, pragmatic, etc.. Speech in which prosodic knowledge carries no information would be analogous to reading a phonetic transcription of a sentence in which punctuation mark, spaces between words and stress markers had been removed (O'Shaughnessy, 1976). In such situations all of the needed information is there for most of the sentences but it certainly would be simpler if the delimiters and other markers are left in. This analogy may not be adequate enough to describe the importance of prosodic knowledge in continuous speech because prosody carries much more information than the punctuation marks.

There are several advantages for spoken messages over written version. For example, sentences can assume different meaning depending on how the speaker pronounces it. To convey a special meaning of a sentence the writer is limited to tools such as italics and punctuation marks, which are frequently insufficient. A speaker can modify his utterance by changing prosodic parameters with a related change in the meaning that are not explicitly contained in the lexical and syntactic makeup of the equivalent text. Even when the sentence does not have special meanings and the speaker is not expressing emotions, prosodic knowledge provides many useful clues in speech. For example, spaces between words and the use of punctuation mark provide syntactic function in written sentences, whereas in fluent speech there are no significant pauses between most of the words and no obvious way of dividing them up into words and phrases. But it' is possible to syntactically disambiguate the sentences by making use of prosodic knowledge. Acquisition and incorporation of prosodic knowledge will improve the performance of speech systems such as text-to-speech, speech-to-text and speaker recognition systems significantly.

In this thesis we study the properties of intonation knowledge for continuous speech in Hindi, an Indian language and its application to various

speech systems. Intonation is manifested as the variation of Fo with time. An utterance may convey a different meaning by changing the intonation pattern even if it composed of the same segmental phonemes. It also helps to group words into syntactic blocks for the semantic interpretation of the utterance. For example, emphasis of words in an utterance is accompanied by significant fall and rise patterns of the corresponding pitch contour.

1.2 Significance of intonation knowledge in speech systems

The function of a text-to-speech system is to convert an input text into a speech output. It finds widespread applications in man-machine communication. Most of the existing text-to-speech systems typically suffer from improper duration and lack of intonation knowledge. In a text-to-speech system intonation refers to the periodicity of the source for voiced speech sounds. In order to incorporate intonation knowledge in a text-to-speech system one need to represent features of F_0 contour in some representable form. Intelligibility and naturalness of the synthetic speech improve significantly due to the addition of this knowledge.

A speech-to-text system converts input speech into a sequence of physical symbols and these symbols are later manipulated into meaningful text. Continuous speech contains information regarding with acoustic phonetic behavior of speech sounds and several higher level knowledge sources related to syntax, semantics, pragmatic, etc.. Intonation is one of the manifestation of these higher level knowledge sources in speech. By exploiting the properties of intonation patterns, it is possible to hypothesize some word boundaries from continuous speech, to classify words into content words and function words and to determine syntactic properties of the sentences such as type of the sentence, location of syntactic boundary, etc.. Also, intonation knowledge is helpful for disambiguating speech signals into symbols at different stages of acoustic-phonetic decoding.

Based on the differences in voice characteristics, the acoustic properties of same text may vary from speaker to speaker. Automatic speaker recognition is a

process to extract the features of these speaker variabilities from the speech signal. Properties of intonation pattern and other prosodic features are helpful for identifying people from their voices. Such properties are classified as learned features or time variant features. Since these features change with respect to text, they are suitable for text dependent speaker recognition. By using the properties of intonation patterns we can identify features which are defined selectively for certain speech events at appropriately chosen locations in the utterance. This can help to reduce the redundancy in data to a large extent. Prosodic features can be successfully used for speaker recognition even if speakers have similar vocal tract characteristics. But it is relatively easy to mimic when any one of these features are used in isolation.

1.3 Scope of the work

There exist several problems which make speech processing difficult and unsolved at the present time. They are summarized as follows: (1) Like spaces in written language there is no separator or silence between words in continuous speech. (2) Each basic unit is modified by its preceding and following unit. Also properties of sound units are varied in larger contexts like its place in the whole sentence. (3) Features of speech signals vary significantly due to intraspeaker variabilities like speaking mode (singing, shouting, whispering, stuttering, with a cold, creakiness, voice under stress, etc.), interspeaker variabilities (different timbre, male, female, child, etc.) and due to the signal input device (type of microphone) or due to the environment (noise, cochannel interference, etc.). (4) Same speech signal carries different types of information such as syntactic structure, meaning, sex, identity of the speaker, his mood, etc. A system has to focus on the kind of information which is of interest for its task. (5) There is no precise rule at present for formalizing the information at different levels of decoding (including syntax, semantics and pragmatic). This makes analysis of continuous speech very difficult. Moreover, these different levels seem to be heavily linked each other (syntax and semantics, for example) and different types of information will cooperate with each other to make the speech signal

understandable, despite the ambiguity and noise that is present at each level (Mariani, 1989).

In this context, we confine our studies to declarative and interrogative sentences in Hindi. Other types of sentences were not considered in this research study. While collecting the data, we instructed the speakers to read the text in a neutral phonetic context to avoid the effects of semantics and emotions. Speaking rate was not changed significantly from speaker to speaker. Also the data was collected only from adult male native speakers of Hindi. We did not try with female and child speakers. The influence of semantic factors are largely ignored throughout the study. For studying the effect of segmental factors on Fo contour we did several controlled experiments in which a nonsense test word was placed in a carrier sentence. These controlled experiments put severe limitations to the analysis because the prosodic properties of the spontaneous speech change significantly due to various semantic features, rate of speech, attitude towards what is being spoken, sex and age of speaker, etc.. But studying the prosodic properties from spontaneous speech is very difficult and we cannot make any valid generalization of the properties in such cases. Also, the major motivation behind this study is to use these properties in the development of speech systems. The objective of our text-to-speech system is intended to perform the function of a reading machine and the speech-to-text system is to perform the function of a dictation taking machine. The development of our speaker recognition system is based on a fixed text. For all these cases semantic and other aspects of spontaneous speech play limited role.

1.4 Organization of the thesis

The main focus in this thesis is on the acquisition of intonation knowledge from continuous speech in Hindi and on the incorporation of intonation knowledge in speech systems for Hindi. This thesis is organized as follows:

In Chapter **2** we review the contemporary work on intonation in different languages. Section 2.2 reviews the properties of intonation patterns in different foreign languages such as English, French, German, Dutch, Japanese, Danish,

etc. and related work in Indian languages. Section **2.3** reviews the issues in the incorporation of intonation knowledge in various speech systems for different languages. Section **2.4** briefly outlines our approach to the problem.

We have used two methods for extracting the pitch contour. One is based on simplified inverse filter method and'the other is based on the properties of group delay functions. These methods are discussed in Section **3.2.** Section **3.3** discusses the general properties of intonation patterns in simple declarative sentences and interrogative sentences in Hindi. The section includes discussion on the declination/rising tendency and the local fall-rise patterns in Hindi.

Complex sentences in Hindi are associated with resetting of Fo contour at major syntactic boundaries. The issues related to resetting of Fo contour are discussed in Chapter **4.** Section **4.3** discusses different physiological, syntactic and semantic factors which can affect F_0 resetting. F_0 resetting is accompanied by a significant amount of pause. Factors which control the amount of pause in continuous speech are studied in Section 4.4.

The influence of segmental factors on F_0 contour is less when compared with the above properties of intonation. In Chapter 5, we discuss the segmental constraints on Fo contour. Section 5.2 deals briefly with a study of the acoustic-phonetic properties of speech sounds. Inherent F_0 of vowels and the effects of adjacent consonants and syllables on F_0 are discussed in Section 5.4.

Our study on intonation is mainly motivated by its significance in an unrestricted text-to-speech system for Indian languages. Chapter **6** discusses the significance of intonation knowledge in a text-to-speech system for Hindi. We are developing a text-to-speech system for Hindi based on parameter concatenation model. Section **6.2** discusses the major design issues in the development of the system. Issues in the incorporation of intonation knowledge in our system are discussed in Section **6.3.** The results from the evaluation of quality of synthetic speech are discussed in Section 6.4.

In Chapter 7 we discuss the applications of intonation knowledge in a speech-to-text system for Hindi. Intonation knowledge is useful for the hypothesization of word boundaries from continuous speech which are applicable

in the context of a speech-to-text system for Hindi. For this purpose we derive pitch accent features from the local fall-rise patterns for continuous speech in Hindi. Section **7.2** discusses pitch accent features in Hindi. An algorithm for the hypothesization of word boundaries is discussed in Section **7.3.** Word boundary hypothesization algorithm is useful for the hypothesization of function words from continuous speech in Hindi. In Section **7.4** we study an algorithm to hypothesize function words from continuous speech in Hindi.

Chapter **8** discusses the design and development of a speaker recognition system based on the intonation knowledge. Design of a speaker recognition system involves two parts, one corresponds to feature extraction and the other corresponds to classification. The features for our speaker recognition system are derived from the pitch accent patterns using the word boundary hypothesization algorithm. Section **8.2** discusses the issues related to the feature extraction of our speaker recognition system. We use neural networks based architectures for identification of the speakers from these features. In Section **8.3** we discuss the architectures of neural networks used for our speaker recognition system. The results of the speaker recognition experiments are discussed in Section **8.4.**

Chapter **9** summarizes the work described in the thesis. Section **9.2** gives the major contributions of the work. In-Section **9.3** we discuss the limitations of our work and in Section **9.4** we discuss the directions for future work.

Chapter 2
INTONATION KNOWLEDGE - A REVIEW

2.1 Introduction

The significance of prosodic knowledge at various levels of speech processing has been discussed in several studies. These studies range from the acoustic-phonetic level to the semantic-emotional level. Intonation knowledge is an important prosodic knowledge which can be used in the context of various speech processing systems, such as text-to-speech, speech-to-text and speaker recognition systems. In what follows, we shall review some of the relevant work on intonation knowledge.

Physiologically intonation can be explained as the rate of vibration of vocal folds during speech production The presence or absence of vibration of the vocal cords are mainly due to the tension of vocal folds and the air pressure across the glottis (subglottal air pressure). Many studies have been conducted in the past to evaluate the effect of each of these factors on the production of Fo (e.g., Lieberman, 1967; Vanden Berg, 1968; Collier, 1975; Atkinson, 1978; Cooper & Sorensen, 1981; Titze, 1989; etc.). The importance of vocal fold tension and subglottal air pressure are summarized in the following paragraphs.

Fo of continuous speech increases with the tension of the vocal folds. This can be done by either the lengthening of vocal folds which results in stretching of effective vibrating portion or by contracting the vocalic muscles. Lengthening of vocal folds causes the increasing of the distance between the points of attachment (cricoid cartilage and arytenoid cartilage) by means of cricothyroid muscles. The stretching of these muscles increases their extrinsic longitudinal tension.

Contraction of vocalic muscle (thyro-arytenoid muscle) reduces the extrinsic tension of vocal folds. But it produces an intrinsic tension due to the isometric contraction of this muscle provided the distance between points of attachment is kept constant. The rate of vibration of vocal folds reduces with the reduction in intrinsic or extrinsic tension. This reduction is primarily due to the relaxation of these muscles after a rise in frequency ('t Hart, Collier & Cohen, 1990).

Like the tension of vocal folds, pressure below the laryngeal region has also been useful in characterizing the production of F_0. This pressure is called subglottal air pressure. It is controlled by respiratory muscles. The increase in subglottal air pressure increases the amplitude of vibrations of vocal folds by forcing them to widen apart. This leads to a mechanical stiffness which in turn causes a higher value of F_0 because stiffened cords bounce back faster. However, the effect of subglottal air pressure is small compared to the effect of vibrations of vocal folds on the production of F_0. The slow and gradual decrease of subglottal air pressure over the course of an utterance results in overall reduction of Fo. It is called declination of F_0 contour.

The F_0 contour is also affected by the articulatory gestures of the speakers, such as jaw lowering, raise of tongue body, elevation of soft palate, etc.. These effects are not intended by the speaker and can be described using the articulatory properties of the speech sounds. These cannot be considered as constituent of the Fo contour as a linguistic entity and can be classified as the segmental factors on the production of F_0.

Properties of intonation patterns of speech vary with the language. A considerable amount of studies have been done for exploring the properties of intonation in various languages. In this chapter we review the properties of intonation patterns of speech correspond to different languages and their applications in various speech systems. This chapter is organized as follows: In Section 2.2 we review the studies on the properties of intonation patterns. It includes a discussion on the properties of intonation patterns in foreign as well as Indian languages. Applications of intonation knowledge in speech systems are reviewed in Section 2.3. Section 2.4 outlines our approach to the problem.

2.2 Review of studies on intonation knowiedge

The linguistic factors such as syntax, semantics, morphologic, phonologic and pragmatic constraints of a language change intonation patterns of the language significantly. But all speakers are equipped with similar production and perception apparatus, and consequently have similar capabilities and face some similar physiological constraints. Due to the reflection of these similarities, some properties of intonation patterns are considered as language independent (Vaissiere, 1983). This section discusses both language dependent and language independent properties of intonation patterns. It is divided into two subsections: studies on the properties of intonation patterns in foreign languages and of Indian languages. The reported work in latter section is very little and there is hardly any literature which discusses the significance of intonation knowledge in the context of speech systems for Indian languages.

2.2.1 Studies on the properties of intonation patterns of foreign languages

Properties of intonation patterns can be divided into global and local. The global characteristics of F_0 can be defined over a unit of speech that includes an entire class or phrase. Declination tendency and rising tendency are the salient global attributes in declarative sentence and interrogative sentence, respectively. Local attributes represent those characteristics which include no more than two adjacent words. Fall-rise pattern of F_0 and the effects of lexical stress are considered as local attributes (Cooper & Sorensen, 1981). It is difficult to partition the properties of F_0 contour into global and local for a given utterance. In this section we discuss some major properties of F_0 contour reported in literature. The discussion includes the global properties of intonation patterns, local fall-rise patterns, resetting of F_0 contour and the effects of segmental factors on F_0 contour.

2.2.1.1 Global properties of intonation patterns

Declination tendency and rising tendency are characterized as global attributes in declarative sentence and interrogative sentence, respectively. The

form and domain of global properties of F_0 contour, especially declination were investigated for large variety of sentences in several languages. The basic aim of this study involves the formulation of a quantitative model of declination that is sufficiently abstract to capture the essence of global attributes as it commonly appears in declarative sentences. Global properties of F_0 contour are more perceivable in speech than any other properties and hence a detailed analysis of this attribute is required as a foundation for studying other features of Fo contour, such as local fall-rise patterns, resetting of F_0 contour and segmental factors. In the following paragraphs we discuss some of the characteristic features of declination/rising of F_0 contour for different languages.

The characteristics of declination of F_0 contour were investigated for large variety of declarative sentences in English. In his famous study, Lieberman (1967) explained declination of F_0 contour as a universal phenomena of human speech except for some predictable classes. The physiological basis of this phenomena may be condition of least articulatory control. If the speaker does not increase the tension of vocal folds deliberately to counter the fall in subglottal air pressure during phonation, the F_0 falls accordingly. This pattern of articulatory activity thus produces an intonation pattern which can be used to delimit unemphatic declarative sentences in normal speech. He defined this pattern of articulatory activity as *archetypal normal breathgroup*.

In 1969 Majewsky and Blasdell investigated F_0 cues for Polish and American listeners to make statement and question decision based on declination/rising tendency. According to them, for each speaker there is a substantial difference between statement and question, not only in the rise or fall of the final segment, but the initial segments of the contour as well. The results of this study indicated that the most prominent cue for a statement-question identification is the absolute value of end point frequency. After studying F_0 contours for simple declarative sentences in English, Olive (1975) concluded that the parameters, such as amplitude, duration and formant coarticulation required only minimal change to make the speech sound continuous if a meaningful F_0 contour is imposed on a sentence. The results of his investigation suggested that

a pattern in F_0 contour can be expressed as a function of grammatical phrase structure of the language. The first and the last meaningful words in a sentence are quite different from each other and from the rest of the words in the sentence. Olive showed that there exists a difference in the slope of the Fo contour for a final and a nonfinal word in a phrase which might help to communicate the phrase boundaries.

O'Shaughnessy (1976) schematized declination of F_0 as a single line with negative slope drawn near or through low values of F_0 occurring in an utterance. This reference line is termed as base line and local fall-rise patterns are defined as the height above the base line. According to O'Shaughnessy the overall pattern of F_0 in a breathgroup for English utterances other than yes-no type questions starts on a relatively low level, rises rapidly on the first emphasized syllable, and then gradually falls to a very low level. In yes-no type questions, Fo differs from the typical pattern by falling left after the initial rise and rising rapidly to the highest level at the very end in the utterance. Wh-question in English is characterized by emphasizing a particular word in the sentence. Later in an attempt to specify the F_0 contours of speech from linguistic specifications, O'Shaughnessy and Allen (1983) stated that a typical F_0 contour for an English utterance can be viewed as a composite result of a set of hierarchical patterns associated with the sentence, phrase, word and phoneme, with the Fo effects of such successive (more local) pattern being superimposed upon the previous higher (more global)F_0 phenomena. For instance, if a sentence contains more than one phrase, the Fo properties of each phrase is superimposed on the Fo contour of the sentence and so on.

Many experiments were carried out by Pierrehumbert (1979) to investigate how declination of F_0 contour in English speech is perceived by listeners. From her studies we can conclude that speakers normalize for declination in judging the relative heights of peaks of intonation contours. This implies that the relative salience of stressed syllables in neutral intonation is not directly mirrored in the values of F_0. Pierrehumbert observed that declination of F_0 contour carries the information about the syntax of the sentence and hence could be used for one

level of linguistic processing. Her experiments indicated that the listeners expect more declination in wide pitch range than in narrow pitch range and the expected slope of declination is less for a longer utterance than for a shorter one. The amplitude down drift which is typically accompanied by down drift of Fo contour was also found to have a part in mental representation of declination. In a later work, Pierrehumbert (1981) defined declination as an implicit level in computing Fo levels which corresponds to the different degrees of prominence. This observation was supported by various perceptual experiments. For instance, in a sentence for the speech to sound equally stressed at any two instances, one generally observes a higher F_0 at the first instance as compared to the second. On the contrary, when two instances have the same F_0, the sound at the second instance was perceived as more stressed. Based on these observations Pierrehumbert developed a schematized model for intonation.

One of the important studies on the properties of English intonation was done by Cooper and Sorensen (1981). This study was motivated by the inability to predict the salient characteristics of F_0 contour using base line model proposed by O'Shaughnessy (1976). Moreover, it is not possible to define the local fall-rise patterns accurately by giving any particular base line. Cooper and Sorensen conducted a detailed investigation on the properties of top line, which drawn through the successive peak values of F_0 contour. Top line was selected for analysis because: (1) it strongly depends on speaker's coding of linguistic structures, (2) it is relatively easy to measure and (3) it is perceptually more important to the listener (Cooper & Sorensen, 1981). They concluded that the form of declination can be captured by an empirical formula which adequately predicts the value of intermediate peaks of Fo contour throughout the course of a large variety of single clause declarative sentence structures in English. They called this as top line rule. To develop the top line rule, they hypothesized that the peak of Fo contour of the first stressed syllable in a given sentence is higher for the longer main clauses. O'Shaughnessy (1976) also made a similar observation. Cooper and Sorensen proved that the top line rule is approximately invariant across different situations like sex of the speaker, length of clause, type

of clause, type of grammatical category, rate of speech, etc.. Since this rule predicted the peaks of F_0 contours in utterances at different situations, they concluded that the top line rule captured the declination attribute. Even though the top line rule provided a good characterization of declination of Fo contour in declarative sentences in English, it can not be considered as a linguistic universal. Cooper and Sorensen have also shown that it is not possible to capture the form of declination of Japanese sentences even though both English and Japanese exhibit general features of declination. This was because Japanese speakers appear to program their values of F_0 over shorter domains than English, and the inherent pitch accent of individual words play a much greater role in values of Fo in Japanese speech.

In 1988 a model for intonation in British English was suggested by Willems, Collier and 't Hart, They revealed the existence of declination even for spontaneous speech and modeled declination of F_0 by drawing a base line through the lower stretches of F_0. The main aim of their investigation was to map perceptually relevant features in British English melody in terms of pitch movements. According to them, pitch movements could differ one another along various dimensions, such as direction, range and duration, and timing with respect to the syllable structure. A stylization procedure was used to generate a set of perceptually relevant movements for any F_0 contour. Later, 't Hart, Collier and Cohen (1990) mentioned the use of three parallel lines (top line, base line and a middle line) to model F_0 declination in British English.

Although most of the studies on intonation is centered around English, various noticeable studies on intonation is emerged from other languages, such as Dutch, French, Danish, Japanese, German and some tone languages. In the following sections we review the studies on the global properties of intonation patterns in some of the above languages.

't Hart, Collier and Cohen (1990) had made an extensive study on the properties of intonation patterns in Dutch. In one of the earlier studies Cohen and 't Hart (1967) divided the properties of intonation patterns in Dutch into major class, minor class and micro class. Properties of the major class correspond

to the global properties of intonation. Cohen and 't Hart stated that the pattern for major class found to resemble a hat which consists of an initial gradual fall-off, a steep rise, an upward shifted segment of declination line, a steep fall and a final declination which is the extension of the initial declination. The positions of rise and fall coincide with prominent syllables of these words that play a dominant part in the utterance. In a later study 't Hart, Collier and Cohen **(1990)** analysed the effects of declination in various levels such as acoustic, perceptual, production and communicative aspects. They concluded that the effect of declination can not be brought in a satisfactory way by merely applying smaller rises and falls; the tilted line should be present as such. They also followed the assumption of base line proposed for English and considered it is more difficult to draw one line through the peaks than through the valleys. To stylize the properties of intonation patterns in Dutch, 't Hart, Collier and Cohen used a top line which is parallel to base line. They suggested that the parallel top line and base line are not judged less natural than those with convergence. Terken and Lameer **(1988)** had also conducted studies on the perceptual quality of intonation patterns in Dutch and underlined the need of natural intonation for the better intelligibility of speech.

Gussenhoven and Rietveld **(1988)** also considered the effect of declination of F_0 in Dutch. They tested three hypotheses to explain effect of declination. They are: **(1)** time dependent uniform sloping down of the contour, (2) lowering of the final portion of the contour and (3) time dependent peak-to-peak lowering function. Based on the experiments, they found the declination effect is due to time dependent down sloping and final lowering, and peak-to-peak lowering function did not have any effect on intonation. Rietveld and Gussenhoven **(1987)** conducted another study to analyze the effect of rate of speech on the properties of intonation patterns in Dutch. Results of this study suggested that listeners are biased towards the temporal features that regularly accompany the intonation structures concerned. According to them, the presence of high and fairly monotonous section in the intonation contours leads to an increase in rate of speech in perception.

Vaissiere (1974) described the F_0 contours of French utterances as a sequence of simple patterns super imposed on a falling base line. Long utterances were divided into breathgroups and each breathgroup was characterized by a separate base line. Later studies on French intonation was mainly based on this frame work. Delgutte (1978) studied the effects of syntax and attitude of the speaker on F_0 contours in French. The main influence of syntax is expressed by different prosodic boundaries occur at the major syntactic boundaries. Delgutte also proved that the attitude of the speaker influences the global properties of F_0 contour significantly. In English this was proved by Williams and Stevens (1972) much earlier.

Thorsen (1980) analysed the properties of intonation patterns in different types of Danish sentences, such as declarative, nonfinal (incomplete) and interrogative sentences. Thorsen used a standard base line to study the effects of declination of F_0 contour in Danish speech and observed that a steeply falling intonation contour was associated with a declarative sentence. In interrogative sentences, intonation contour was least falling and the slope of declination for a nonfinal sentence was between declarative sentence and interrogative sentence. The changes in word order change the slope of declination of F_0 contour. For instance, the slope of declination of interrogative sentence could be increased further by inverting the word order. The global properties of intonation patterns of interrogative sentences in Danish proposed by Thorsen contradicted with several other languages which assumes a rising tendency for interrogative sentences. Thorsen also studied differences between declarative and interrogative sentences with contrast emphasis, the effects of the lowering of F_0 of sentences in a text and the differences in the slope of sentence intonation were induced by different boundary conditions (Thorsen, 1983; Thorsen, 1985). Based on the experiments, she suggested that these issues can be handled by a linear sequence model with more parameters to determine the pitch accents. In a later addition, Thorsen (1986) suggested that the F_0 lowering of Danish sentence is more related to semantics than syntax.

Fujisaki and Hirose (1984) made a study on F_0 contours of declarative

sentences of Japanese. The proposed model for generating Fo contours of spoken sentences in Japanese elucidated the relationship between F_0 contours of the sentences and the linguistic and nonlinguistic information. It was based on a quantitative formulation of the process where by logarithmic F_0 was controlled in proportional to the sum of two components correspond to the effect of phrase and accent. According to them, the shape of phrase component was responsible for the natural declination of F_0 contour and it did not fluctuate appreciably with the sentence length. This model successfully reproduced the essential characteristics of F_0 contours of declarative sentences in Japanese.

Ladd (1983) compared the intonation properties of German with several available models of intonation. He found F_0 contours of interrogative sentences exhibit declination in addition to rising accents. In general, pitch rises from the accented syllable and gradually declined from its peak until it steps down to the next accented syllable. He outlined a peak feature representation of intonation and compared with other current models. Ladd claimed that his model of declination accounts for phonetic data better than models involving overall tunes or base line and top line. In one of the later work Ladd, Silverman, Tolkrnitt, Bergmann and Scherer (1985) analysed the F_0 contour for German and showed that ranges of F_0 contour had independent effects due to interspeaker differences and differences in verbal contexts.

Intonation behavior of some tonal languages were also supporting the declination property (Lindau, 1986). Lindau separated intonation patterns as sloping grids of parallel lines, inside which tones are placed. The tones are associated with the turning points of F_0 contour.

There exist some exceptions for the declination property of declarative sentences. For instance, the model of F_0 contour generated by Garding (1983) for Swedish and Greek did not support the arguments for declination. Also in another study Umeda (1982) proposed that declination of Fo contour is not a global behavior but a situation dependent phenomena. But the arguments against declination of F_0 contour is very few when compared with its supporters.

2.2.1.2 Local fall-rise patterns

Global properties of F_0 contours are a backdrop property to which local fall-rise patterns are superimposed. Properties of local fall-rise patterns are different for different languages, and even for the same language the interpretation of this properties change with the method of analysis. The study on fall-rise patterns are so divergent because there are no linguistic universal like the declination of F_0 contour. In the following paragraphs we summarize the important studies on local fall-rise patterns of Fo contour for different languages.

While discussing acoustic correlates of emotions on speech Williams and Stevens (1972) pointed out that the normal F_0 contour is characterized by smooth, slow and continuous changes of F_0 as a function of time and the changes occur in syllables on which emphasis or linguistic stress is to be placed. Olive (1975) discussed the changes in F_0 contour as the effects of phonetic properties of the words and the influences dictated by the grammatical structure of the sentence. He considered the variation of F_0 contour with respect to the semantic properties also.

Hierarchically local fall-rise patterns are associated with phrases and words in sentences. Hence stress patterns of words in continuous speech are different from each other. According to O'Shaughnessy and Allen (1983) emphasis is one of the major factor to determine the stress pattern. It is signaled by significant rise in Fo contour. In the case of little or nonemphasized words, any rise in Fo contour emphasizes the syllable on which it starts. Also, a sharp fall of Fo occurs immediately after the vowel of the emphasized syllable. For English these fall-rise patterns were initially suggested by Lehiste and Peterson (1961).

O'Shaughnessy and Allen (1983) suggested the F_0 contour of a syllable occurs between two F_0 rises have a less F_0 fall compared with others. Since Fo pattern for a declarative sentence in English is characterized by two declination lines (top line and base line) which converges towards the end, the successive F_0 rises decrease so that emphasis rises are smaller towards the end of the utterance. So as per O'Shaughnessy and Allen, F_0 emphasis in two utterance should be

compared by sentential position. Pierrehumbert (1979) also pointed out that the changes in F_0 contour was significant at the beginning of an intonation phrase than at the end.

According of Pierrehumbert, local fall-rise patterns in pitch contour carry information about the relative salience of different syllables. She considered a series of target values to determine the local fall-rise patterns of Fo contour (Pierrehumbert, 1981). They were: (1) high target values at the onset, (2) a low target value on the first stressed syllable and (3) a high target value on the second stressed syllable. Pierrehumbert decomposed phrasal tunes in English into targets associated with margins of phrases. The tonal marking of a stressed syllable may be either single target or a sequence of two targets. The transition between two targets were not always monotonic. If two targets are high and sufficiently separated in time, F_0 shows a sagging behavior in between.

Cooper and Sorensen (1977) studied the effect of local fall-rise patterns of F_0 contour and opposed the arguments about the influence of phonetic conditioning factors of stress. According to their opinion, fall-rise patterns are properly attributed to speaker's syntactic code. Fall-rise patterns of F_0 contour is accompanied by the boundaries of clauses and the major phrases in a variety of sentence types in English. Cooper and Sorensen (1981) suggested that the magnitude of fall-rise pattern is proportional to the strength of the associated boundary. Typically fall-rise patterns are superimposed over not more than two words and it is different from the global properties of F_0 contour which is manifested at a longer stretch of speech. According to them, the type of phrase structure representation adopted by linguistic grounds provides a good first approximation to the type of internal syntactic representation computed by the speaker. Also, fall-rise patterns are separately influenced by left and right side of a given syntactic boundary. Cooper, Soares, Ham and Damon (1983) studied the effect of emphatic stress on local and global properties of F_0 contour. Results of their studies proved that emphatic stress was accompanied by a large increase in the peak of Fo contour of the emphasized word and by a small but systematic decrease in the peak of F_0 contour of the following word. Hence increase in

emphatic stress could be attributed to an increase in the tension of vocal folds. Also they showed that the increase in the rate of speaking produces slightly higher values of F_0 than normal.

Holmes (1988) also underlined the fact that English shows a general tendency for pitch to fall gradually from the beginning to end with many local variations. For polysyllabic words, one syllable has the primary stress with main pitch movement and the other syllables of the word are either unstressed or carry less prominent secondary stress. Holmes called the syllable with the biggest pitch movement as nuclear syllable and the pitch pattern associated with nuclear syllable as nuclear tone. Nuclear tone for a simple declarative sentence is pitch fall and yes-no type question is a substantial pitch rise.

Collier (1975) suggested that the local fall-rise patterns of F_0 contour are caused primarily by the action of cricothyroid muscle. According to him, the increasing activity of this muscle raises F_0, its continued contraction keeps F_0 high and the lowering of F_0 was due to its relaxation. Later Willems, Collier and 't Hart (1988) concluded that a pitch contour can not be mapped on to the words that makeup a clause or a sentence unless certain accentual and surface syntactic properties of the text are taken into account. According to their *flexibility* principle, when a contour is mapped into a clause, its pitch movements are flexibly moved to those syllables that require a pitch accent or to those syllables at constituent boundaries that need a continuation cue. In a later work 't Hart, Collier and Cohen (1990) modeled intonation patterns in Dutch as a sequence of pitch rises and falls. These were further differentiated according to variables such as positions in the syllable, size and slope. These pitch movements should allow a unique specification which is suitable for all distinct intonation patterns. Also they should allow for the specifications of all the pitch movements which makeup the pitch contour of an utterance. They considered the pitch movements with a straight forward interpretation at the phonetic level.

While discussing a model on French intonation Vaissiere (1974) observed that the local fall-rise patterns were largely co-occurrent with content words (semantically meaningful words) of three syllables or more. Shorter words

receive an incomplete pattern or share the pattern with the preceding or following word. Delgutte (1978) successfully attempted to generate F_0 contours for French sentences by superimposing above patterns on a base line. He found that patterns of some words are low and adjacent to base line. He called such patterns as null patterns. Also, he introduced another type of patterns called emphatic pattern which was characterized by a sharp rise associated with a peak on the first syllable and lowering on the last syllable of a word. Delgutte's model was unable to predict the patterns of monosyllabic words.

Study on the properties of intonation patterns in Danish by Thorsen (1980) suggested that local fall-rise patterns have the same basic shape with relatively low stressed syllable followed by a high falling tail of unstressed syllables. The magnitude of rise of F_0 from stressed to post-tonic syllable varies with time and sentence type. In a later study, Thorsen (1983) discussed the effect of emphasis also. The stressed syllables of emphasized words will stand out clearly from the surroundings. This is brought about by a rising of F_0 (except initial position), an elaborate rise within that syllable and a deletion of deflections of F_0 in neighboring stress groups. The immediately surrounding syllables fall away sharply from the stressed syllable of the emphasized word. Thorsen was also admitting the existence of a varying degree of prominence among stressed syllables of an utterance without necessarily evoking impression of emphasis for contrast.

Even though fall-rise patterns are superimposed over the global properties of Fo contour, each analysis interpreted them differently. By this study we can conclude that the fall-rise patterns are syntactically dictated and can be determined by the phonological pattern of the language. Hence they are considered as language dependent.

2.2.1.3 Resetting of F_0 contour

Declination of F_0 contour gets modified across major syntactic boundaries. This is imposed by both the physiological factors of speech production mechanism and syntactic properties of the language. As we discussed earlier,

subglottal air pressure drops during phonation. Speakers are forced to give a significant amount of pause between the utterances to compensate drop in subglottal air pressure. During a pause, subglottal air pressure is built-up again which is manifested as resetting of F_0 contour. In the following paragraphs we review the issues related to the resetting of F_0 contour for different languages.

According to Lieberman (1967), it is physiologically more convenient to divide a sentence into more than one breathgroup. Each breathgroup corresponds to a phonemic phrase. An ambiguous sentence can be disambiguated into constituent structures in terms of different phonemic clauses. Lieberman proved this for several sentences in English. Streeter (1978) also observed that speakers can modulate F_0 contour to group words together which constitute a major syntactic unit and hence effectively used F_0 contour for disambiguating long sentences into phrases of smaller size.

O'Shaughnessy (1976) observed the presence of major syntactic breaks at certain positions of utterances. These breaks are indicated by discontinuities in Fo contour such as a sharp fall of F_0 followed by a rise on the final voiced phone before break. Also he observed large continuation rises at major breaks between clauses which often cause the falling pattern to reset at higher level, from which it resumes falling after the break.

Cooper and Sorensen (1977) studied the properties of resetting of F_0 contour with respect to the syntactic knowledge of the language. They observed that F_0 contour resets to new starting value at the boundary between two main clauses if the sentence contains more than one main clause. The slopes of declination lines for different main clauses are equal. In a later study, Cooper and Sorensen (1981) concluded that the entire utterance exhibits a domain of declination of their own and resetting of F_0 contour seems to be triggered directly by the presence of syntactic boundaries, and did not depend on accompanying breath pauses. Even though they analysed the effect of resetting with respect to syntactic structure, they suspected that a variety of nonsyntactic factors such as long clauses, slower rates of speech and lower semantic relatedness between two clauses also influence the resetting of F_0 contour.

Based on the experiments, Ladd (1988) suggested that hierarchical structure of the sentences has an effect on the amount of resetting of F_0 contour. If the boundary is stronger, following peak of F_0 contour seems to be higher. Also, he observed partial reset in certain cases which he called as ***declination within declination.***

Studies on F_0 contours of Dutch by 't Hart, Collier and Cohen (1990) suggested various aspects of resetting of F_0 contour. As per their studies, making a reset did not require inhalation as a means to increase the subglottal pressure. They also noticed that taking a new breath does not necessarily cause the F_0 to rise. According to them, resets are under laryngeal control. While discussing the communicative aspects of declination, they mentioned that Fo of a speaker is bound to a lowest limit. When declination goes below this, it is interrupted by a reset. In other words, if declination of F_0 is nil, resetting would not be necessary.

One common acoustic feature accompany with resetting is the presence of pause. But the relation of pause with respect to resetting is explored very little. According to Lieberman (1967), pauses are made for the purpose of taking breath and to make the meanings of the word clearer. Positions of the pauses are determined based on the semantic interpretation of the utterances. But as per Kutic, Cooper and Boyce (1983), pauses are determined by the syntactic factors of the sentences. They serve to clarify the subgrouping of smaller units. Their experiments confirm the presence of pauses at the boundaries of parenthetical and main clauses. They also noticed that the amount of pause before breathing, if any, will be more. At slow and normal rates, speakers tailor their needs to breath to syntactically defined pause patterns.

Based on these studies, we can conclude that resetting of F_0 contour is syntactically determined. The effect of resetting is proportional to the strength of the following syntactic clause. Moreover, the resetting of F_0 contour is dictated by the speech production mechanism to compensate the drop in subglottal air pressure. A significant amount of pause is accompanied with resetting of Fo contour. The amount of pause can also be determined by the strength of the syntactic boundary. That is, the longer the pause, the stronger the syntactic

boundary.

2.2.1.4 Segmental factors on F_0 contours

Acoustic properties of individual phonemes in an utterance alter Fo contours at syllable level. These effects are not very significant when compared with other properties, such as global and local properties of F_0 contour. Moreover, the segmental properties are largely language independent because they are imposed by the acoustic-phonetic constraints of speech sounds. In this section we discuss the segmental properties of F_0 contour.

Lehiste and Peterson (1961) were made a detailed analysis on the segmental features of F_0 contour. They proposed that an average F_0 is associated with each syllable nucleus and is called intrinsic F_0 (inherent Fo). The high vowels /*i*/ and /*u*/ are associated with the highest inherent Fo and the low vowel /*a*/ has the lowest inherent F_0. The inherent F_0 of the central vowels /*e*/ and /*o*/ occur approximately in the middle of the F_0 range. According to them, the selection of a particular pitch allophone is conditioned by the segmental quality of the syllable nucleus. Lehiste and Peterson analyzed the effects of the preceding and the following consonants to the present vowel. Based on these observations, they concluded that higher F_0 occurs after voiceless consonant and considerably lower Fo occurs after the voiced consonant. Even though Lehiste and Peterson proved these factors for American English, this is true for other languages as well. Several other studies also supported these observations (e.g., Ewan, 1979; Ohala & Eukel, 1976; Peterson, 1978; Shaddle, 1985).

Rosenberg (1968) studied the effect of averaging of pitch values on the quality of natural vowels. According to him, pitch period of vowel sounds are smoothed significantly when it is embedded in sentences, but the quality of vowels do not change significantly. Haggard, Ambler and Callow (1970) studied the effects of pitch range at the onset of voicing after a period of articulatory closure for a consonant. They proved that a low rising pitch at the onset of voicing indicates a voiced consonant and high falling pitch at the onset of voicing for a voiceless consonant. Ewan (1979) analysed the acoustic coupling of the first

formant frequency with F_0 for nasals and concluded that their relation is not significant. According to him, several physiological factors, such as tongue pull, jaw movement, tongue compression and pharyngeal constriction had an effect on inherent F_0.

Umeda (1981) was against the assumption of inherent F_0. According to her, F_0 of different vowels show a random pattern. Based on a study of fluent speech, she suggested that the initial and the peak values of F_0 contour of the vowel is depending on preceding stressed consonants. She also found that F_0 of a vowel is higher if the preceding consonant is voiceless stop or voiceless fricative. Also in the reading of extended meaningful text, the direction of F_0 movement at the onset of vowel is not a reliable voicing indicator of the preceding consonant. But the studies on isolated or short carrier phrased utterances of test data contradicted the above observations (Umeda, 1981). Based on this study, she suggested that a proper F_0 control of segmental factors in the production of speech would enhance the intelligibility of the consonants. In her experiments, Umeda controlled the text for consonant contexts, but did not control for a range of effects of intonation, such as lexical stress, sentence stress or declination. Hence she failed to observe the inherent properties of Fo contour (Shaddle, 1985).

O'Shaughnessy (1976) also considered the height of the vowel is proportional to its average F_0. He also showed that the inherent F_0 of the vowels vary with the preceding consonant and the effect of the following consonant on present vowel is very small. Ohde (1984) suggested F_0 as an acoustic correlate of stop consonant voicing after studying the properties of voiceless aspirated, voiceless unaspirated and voiced stops. Ohde also proved that F_0 varies as a function of consonant voicing. That is, F_0 is high after voiceless stops and low after voiced stops. Ladd and Silverman (1984) suggested that the global properties of F_0 contour alter the effect of inherent F_0 significantly. Unlike Umeda (1981) they conducted experiments in sentences with controlled phrase position and stress. They found the evidence of inherent F_0 of vowels and the interaction of inherent F_0 with sentence position. Shaddle (1985) also proved

that inherent F_0 occurs in sentence context. Comparing with his results with other studies, Shaddle indicated that both properties of F_0 contour and pitch accent affect the amount of inherent F_0. Shaddle also showed that the inherent F_0 differences decreased significantly in unaccented final positions of the sentences.

Zawadsky and Gibert (1989) considered the changes of F_0 of a vowel with articulator position. They found that the positions of jaw are more closely related to the Fo than the position of tongue. They did not support the correlation of Fo with height of the tongue. Since both tongue and jaw are associated with upward movements for high vowels and down ward movements for low vowels, one could reasonably expect the combinations of both articulators to be more closely related to F_0. The height of the jaw is a good indicator of the height of the vowel. They also showed that positions of jaw are more important than compression of tongue in determining F_0 due to several physiological limitations.

Thus, the review on segmental properties of F_0 contour reveals the following conclusions. Segmental properties are largely language independent and primarily based on acoustic-phonetic analysis of speech sounds. Each vowel exhibit an average F_0 of its own and it is called inherent Fo. The properties of preceding consonant affect inherent F_0 of the vowel significantly. The effect of the following consonant and the surrounding syllables have not been exploited to its full potential.

2.2.2 Studies on the properties of intonation patterns in Hindi.

Most of the studies on intonation in Hindi is stressed in linguistic point of view, which may not help in the context of the development for speech systems for Hindi. In one of the earlier studies in Hindi, Moore (1965) defined intonation as the phonological level in which the prosodic features of pitch, intensity and quantity function significantly. He subdivided intonation in Hindi into three contrastive subsystems corresponding to emphasis, expression and segmentation. Emphasis system highlights a portion of utterance for emphasis and contrast. The function of expressive system is the communication of the

attitudes of the speaker towards what is being spoken. The function of segmentational system is the division of utterances into pieces corresponding more or less to grammatical units of several ranks. Moore considered average pitch of each syllables measured by analytical means which were more susceptible to errors. Even though he was successful in defining the functions of intonation, his analysis mostly concentrated on the semantic aspects of intonation.

Unlike English and many other ***stress-timed*** (stressed syllables occur at regular intervals of time regardless of intervening unstressed syllables) foreign languages, all of the languages in Indian subcontinent are said to be ***syllable-timed*** (syllables are said to occur at regular intervals of time). That is, no words are differentiated solely by stress due to its marginal role in Indian languages (Ohala, 1991; Crystal, 1985). But a number of linguists assumed the existence of stress in Hindi **(e.g.,** Dimshits, 1966; Sharma, 1969; Pandey, 1989). They suggested some algorithms for the placement of stress in Hindi. These algorithms depend on the concept of weight of the syllable, but they do not always agree as to which syllable gets stressed. Moreover, they never mentioned the phonemic correlates of stress.

Through a series of instrumental investigations on stress in Hindi, Ohala (1986) experimented the significance of duration and pitch as the physical correlates of stress in Hindi. Her results did not agree with the algorithms suggested by the above linguists. In a later study, Ohala (1991) questioned the relation of pitch to stress in Hindi. She suggested that Hindi does not have word stress and stress in Hindi is dictated by the pragmatic factors. Earlier Rumyanceva (1988) also reached similar conclusions based on her experiments. According to her, duration is the primary acoustic correlate of stress in Hindi. But her conclusions may not be valid because she did not consider the short and long vowel differences in Hindi.

In short, the literature on prosodic properties in Hindi are mostly considered the aspects of stress in Hindi. Even though Ohala (1991) conducted some studies on the significance of F_0 contours in synthetic speech, she could not arrive at any conclusion about the properties of intonation patterns in Hindi. We

hardly find any other literature which discuss the significance of intonation knowledge in Hindi in the context of the development for speech systems for Hindi.

2.3 Review on the applications of intonation knowledge in speech systems

Intonation knowledge has many potential applications in speech research. There exist a considerable body of literature which explored the impact of intonation in speech communication. In this section we review the applications of intonation knowledge in various speech systems, such as text-to-speech, speech-to-text and speaker recognition systems.

23.1 Intonation knowledge in text-to-speech systems

In recent years the study on intonation knowledge is mainly motivated by its application in speech synthesis. Unlike other speech systems, here we are able to appreciate the accuracy of the intonation model very quickly through perceptual examinations. Incorporation of intonation knowledge in a text-to-speech system involves various issues. This section reviews the issues in the incorporation of intonation knowledge in various text-to-speech systems.

Mattingly **(1968)** was one of the first researcher who attempted to incorporate intonation knowledge in speech synthesis. In his system Fo was calculated syllable by syllable using a set of pitch values which were logarithmically related to F_0. After the computation, these values were converted to F_0 values by a table look-up. Rabiner, Levitt and Rosenberg **(1969)** incorporated the stress patterns of English in speech synthesis by several combinations of increments in duration and F_0 using a paired comparison technique. They found that the quality of synthetic speech was increased by the incorporation of these rules. Later, Levitt and Rabiner **(1971)** analysed the properties of F_0 contour by subdividing the time axis into a series of time windows and approximately the F_0 contour within each window by a set of orthogonal polynomials. This method provided a basis for comparing the effects of short term smoothing of F_0 contours with that of approximating long term

trends.

According to Olive (1975), the F_0 contour of each word in a sentence can be described as a polynomial with five coefficients. They are: (1) initial Fo, (2) maximum F_0, (3) final F_0, (4) slope at the final point and (5) distance of the maximum F_0 from the beginning. Olive hypothesized that the values of these coefficients are different for different words in the sentence being synthesized. These coefficients depend on the type of the word, the position of the word in the phrase and the position of the phrase of the word in the sentence. These rules were simple and showed the necessary flexibility needed in order to generate speech from any complex text. It also provided rules for semantic as well as grammatical differences. Later Young and Fallside (1979) also developed a similar model of intonation. They added one more coefficient related to the distance between the end F_0 and the maximum F_0. A generator of Fo contour builds up a pitch contour for each word in the utterance as a sequence of primitive contours. Pitch contour for each word group consists of a prehead, a head and a nucleus followed by a tail. These word level pitch contours were combined together to get the complete pitch contour for the sentence.

Umeda (1976) discussed the higher level constraints on text-to-speech systems and developed a set of linguistic rules for a speech synthesis system in English to convey the syntactic and semantic knowledge. Allen (1978) also pointed out the importance of linguistic analysis of sentence level structures to incorporate the higher level knowledge sources. According to them, the major characteristics of F_0 contour, time and intensity of a text-to-speech system for English are their word boundary effects, phrase level F_0 contours, and clause level phenomena. Klatt (1976) characterized F_0 contour by a set of target frequencies which were continuous variables in time. These target values were computed for each syllable nucleus based on information concerning the basic intonation contours, special semantic symbols that may present in the representation for a sentence, and the detailed syntactic structure. A phonetic component was used to compute the effects of low level phonetic factors, such as influences of vowel quality and the preceding consonant on F_0 contour.

Maeda (1974) developed a set of intonation rules for a text-to-speech system in English based on the properties of intonation patterns observed by Lieberman (1967), and Cohen and 't Hart (1967). These rules formed the basis for the generation of F_0 contour in Klattalk speech synthesis system (Klatt, 1987). Here, the intonation contour was modeled in terms of impulses and step commands fed into a linear smoothing filter. Maeda also underlined the importance of syntactic and semantic analysis for getting a natural F_0 contour.

Witten (1982) outlined a model of English intonation in which he divided the utterance into different tone groups irrespective of the type of the sentence. He assigned a pitch contour for each tone group after selecting the major stress point as tonic syllable. Pitch contour on a tone group was specified by ten numbers, each represents a specifiable quantity corresponding to continuation from the previous tone group, notional pitch at start, pitch range on whole of pretonic syllable, pitch range on tonic syllable etc.. To realize this in the synthesis scheme, he adorned the text with prosodic markers and placed a tone group boundary at each prosodic mark. He selected the first syllable of the last foot in a tone group as tonic syllable and the local patterns of pitch contours were assigned with respect to that. The global properties of F_0 contour was determined by the type of the sentence.

The MITalk system (Allen, Hunnicutt & Klatt, 1987) used O'Shaughnessy's algorithm (O'Shaughnessy, 1976) to generate the F_0 contour. The algorithm produces two target values of F_0 for each phonetic segment, one to be used at the onset and one as the mid value. The algorithm can be considered as a cascade of two separate systems. The high level system used syntactic information to sketch the contour. It predicted a superimposed F_0 contour by considering the sentence type, clause contour, phrase contour and individual word contour. The low level system used the information generated by the high level system and additional phonemic data to detail the contour. This reflected the effects of phonemics, lexical stress and number of syllables of the words in the utterance. O'Shaughnessy's algorithm was modified slightly to incorporate the intonation for questions which expect detailed answers. It produces a high peak on the

question word, a steeper falling of F_0 contour than in declarative sentences, and a high peak on last accented syllable than in declarative sentences (Allen, Hanicutt & Klatt, 1987).

Pierrehumbert (1981) suggested a model for synthesizing intonation for English sentences in which F_0 contour was described as a series of target values which were connected together by transition rules. Target values were expressed as locations within current pitch range which vary as a function of time. The syllables corresponding to the target values were determined by the stress pattern of the language. Phrase boundaries may also be assigned as targets. Pierrehumbert's model accepts input text as a string of phonemes annotated with durations, phrase boundaries and target values. If two targets are very close, the dip between them is very small, and relatively unaffected by the height of the targets. The F_0 contour between two targets were analytically determined by the parabola fitted between the targets which incorporated the details up to the phonetic level.

Akers and Lennig (1985) compared the algorithms developed by O'Shaughnessy (1976) and Pierrehumbert (1981) for intonation patterns in English. O'Shaughnessy made use of a detailed part of speech hierarchy to determine accent pitch levels and called it as a naturalistic algorithm. On the other hand, Pierrehumbert's algorithm used the difference between function word and content word to determine the F_0 contour and hence it is called as the schematic algorithm. Both the algorithms used declination line to determine the values of F_0 for each frame of the synthetic utterance. Naturalistic algorithm calculates peaks of F_0 contour as the degree of excursion from the declination line using grammatical hierarchy. They also used a large number of rules to reproduce many details of Fo contours observed in natural speech. The schematic algorithm used two declination lines (top line and base line) which determines the possible range of F_0 as a function of time. These declination lines resets at the beginning of each intonation phrase. To compute F_0 target values, schematic algorithm used an accent target where value is between 0 and 1, determined based on the type of the word. Both the algorithms place peaks of F_0

contour in the vowel region. But if an accented syllable begins with a sonarant sequence, the naturalistic algorithm places the peak in the sonarant. In similar cases the schematic algorithm places a peak later in non-nuclear accents than in nuclear accents (Akers and Lennig, 1985). In short, the main difference between two algorithms is their accentual pattern.

Willems, Collier and 't Hart (1988) developed a synthesis scheme for intonation patterns for British English. It was done by a transition network which splits into six .subsets, each represents rules of sequences for the standardized pitch movements. The model consists of phonetic specification of various types of pitch movements and rules that govern their concatenation into contours based on the transition network. Since the model developed by them is less automated, the users have the flexibility to control the speech output on the basis of his/her linguistic intuition. The model avoided the generation of ill-formed contours by preventing the user from violating the rules of grammar. They tried a similar model for intonation in Dutch also. Quene and Kager (1992) suggested a new model to represent the linguistic structure of the input text for using in a text-to-speech system. Instead of syntactic parsing of the text, their algorithm used a dictionary of function words to establish the prosodic structure of the sentences. Phrase and accent locations were derived from this prosodic structure. This knowledge could be successfully used to generate the F_0 contour for sentences. Some errors were obtained in this model in accentuation which were mainly due to the semantic, pragmatic and contextual factors.

Thus there exist various models of F_0 contours for text-to-speech synthesis. Even though these models were able to reproduce acceptable quality of pitch pattern, it did not generally produce naturally sounding speech (Holmes, 1988). The intonation generated by the models usually sound less interesting than would expect from a human talker. Here, the major limitation is the difficulty pf making a sufficiently accurate linguistic analysis of the text and to relate these linguistic representation to the real world knowledge. In order to get more natural quality speech output from a text, one has to implicitly use detailed knowledge of the syntactic and semantic function of each word in the text for

synthesis which seems to be very difficult for the time being.

23.2 Intonation knowledge in speech-to-text systems

As such intonation knowledge may not be directly applicable for a speech-to-text system because by itself intonation does not contribute to speech information. But as a metalevel knowledge, it can assist the acoustic-phonetic module to disambiguate alternative sequences of units at various stages in a speech recognition system. Intonation knowledge together with other prosodic parameters, such as energy and duration can be successfully used in different sub-tasks of speech recognition like hypothesization of word and phrase boundaries, hypothesization of type of the sentence, etc.. But from literature, it is obvious that the properties of intonation knowledge are not exploited to its full potential except a few cases. In the following section, we review the application of intonation knowledge in speech-to-text systems.

Nakatani and Schaffer **(1978)** considered the stress pattern and rhythm of speech signal as prosodic cues for perception of words. A phrase was labeled ambiguous or unambiguous based on whether its stress'pattern allowed only one or two parsing of the phrase into words, respectively. According to them, pitch and amplitude can not be used in isolation for the perception of words. Through a series of experiments Streeter **(1978)** proved F_0 contour and durational patterns were the major acoustic determinants for the perception of phrase boundaries. According to him, segmental or spectral characteristics of utterances are not reliable cues for the location of the phrase boundaries.

Nakagawa and Sakai **(1979)** analysed the importance of F_0 and energy contours for hypothesization of word boundaries in Japanese. They considered significant excursions in the pitch contour to identify likely areas of word boundaries in continuous speech based on the prior knowledge of possible pitch patterns. Using pitch and energy contours they were able to automatically hypothesize word boundaries between **85%** to **95%** of the time. But the number of boundaries hypothesized was **1.2** to **1.7** times than the number of actual boundaries.

In English several studies in this area had been reported by Lea (1980). Lea had presented an algorithm which used major fall-rise patterns in pitch contour as well as major pauses to identify sentence, clause and major phrase boundaries. Based on a study on English prosody, Lea suggested the importance of stress as a primary ingredient to determine the intonation contour of an utterance. He developed an algorithm to hypothesize stressed syllables from F_0 contour and to determine syllable nucleus of high energy. According to Lea, a detailed analysis of stressed syllables in the utterances help to perform various tasks in speech recognition such as distinction between words, localization of important words, condition for application of phonological rules, perception of phonological structures and organization of articulatory units (Lea, 1980). In the beginning of a new clause, F_0 increases to a high value. Also in the beginning of a major syntactic phrase F_0 increases substantially. Lea used these cues to hypothesize the word boundaries from continuous speech. He also suggested some cues for detecting syntactic structure, such as subordination and coordination. In one of the earlier article, Lea, **Medress** and Skinner (1975) suggested a speech understanding strategy guided by prosodic knowledge. This also discussed the importance of prosodic features to break up continuous speech into sentences and phrases and to locate stressed syllables in these phrases. They demonstrated the use of prosodic knowledge for parsing continuous speech for semantic analysis.

Kohler (1983) studied the influence of F_0 and duration in rhythmic and semantic structuring of continuous speech in German. He considered falls of Fo contour in utterances and hypothesized that each utterance receive a fall of F_0 and is separated by pauses. The peak of F_0 contour is raised if the phrase is expanded or by the addition of another phrase to it. He also developed several other rules applicable to F_0 contour of continuous speech in German and used them to hypothesize prosodic boundaries. **Fujisaki** and **Kawai** (1988) analyzed the prosody of spoken Japanese to realize the linguistic information. The influence of various linguistic factors such as lexical word accent, syntactic structure and discourse structure were discussed based on the analysis of F_0

contour. They suggested that prosodic words could be defined by an accent component, while prosodic phrases and clauses could be defined by the presence of a pause and the resetting of F_0 contour or by the addition of a phrase component. They classified prosodic boundaries and syntactic boundaries based on these and proved that a parallel hierarchy exists between prosodic and syntactic units.

Waibel (Waibel, 1986; Waibel, 1987; Waibel, 1988) was successful in exploiting prosodic knowledge sources for continuous speech recognition systems. For word boundary hypothesization from continuous speech Waibel exploited prosodic cues such as temporal cues (syllable durations, ratio of unvoiced segment duration to syllable duration, and voiced segment duration), intensity profiles and likelihood of stressedness (Waibel, 1987). Waibel used the properties of F_0 contours in English to hypothesize two basic tunes correspond to statements and yes-no type questions. The tune corresponding to statement is a general overall falling F_0 contour and the tune for yes-no type question is accompanied by a final fall-rise pattern. Waibel's algorithm gave an average performance of 78% in hypothesizing the tunes from continuous speech.

Wang and Hirschberg (1992) investigated the textual and intonation features of an utterance to predict the location of intonation phrase boundaries in continuous speech. They suggested that the major phrase boundaries tend to be associated with longer pauses, greater tonal changes and more word final lengthening than minor boundaries.

In short, only a few systems use the prosodic information encoded by the speaker to recognize continuous speech. Hence the knowledge we have acquired on the importance of prosody for human speech perception is largely ignored in this area. It may be due to our inability to explain the variations of these prosodic parameters to the finest detail.

2.3.3 Intonation knowledge in speaker recognition systems

Pitch and other prosodic features can be used for recognizing speakers from their voice characteristics. Prosodic knowledge is treated as a learned feature

which corresponds to the speakers individual speaking habits. Past experiments noted that it is easier to mimic average values of a single prosodic parameter used in isolation. This together with the difficulties in finding the speaker dependent prosodic features cause little progress in using prosodic knowledge for speaker recognition.

Atal (1972) studied the properties of pitch contour for entire utterance and concluded that pitch contour bears important speaker dependent speech characteristics. He represented pitch contour by 40 samples spaced uniformly along the utterance. The pitch data was further compressed to a smaller set by **Karhunen-Leove** transformation and parameters obtained were used for speaker identification. The identification decision was based on the Euclidian distance between the test vector and the reference vectors in the transformed space. The speaker corresponding to the reference vector with smallest distance was selected as the speaker of the test utterance.

Chen and Lin (1987) considered four tone structures in the pitch contour of Mandarin speech for text independent speaker identification. They considered slope, mean and elevation of pitch contour of each word in an utterance as features for recognition. They used vocal tract parameters together with pitch contour for improving the accuracy of the speaker recognition. **Kraayeveld, Rietveld** and Van Heuvan (1991) also considered different features of pitch contours for recognizing the speaker. The features they considered are the Fo at four measurement points, timing at each measurement point, Fo difference between starting point and end point of utterance, pitch rise, pitch fall and their corresponding slopes, and duration of the utterance.

After studying Swedish, English and French Fant, **Kruckenberg** and Nord (1991) concluded that speaker variabilities are language specific. According to them, prosodic aspects of speech contribute more to characterize a speaker than his segmental variations. The rate of prosodic variability is large and more difficultto structure than segmental variabilities.

Thus the properties of pitch contours can be effectively used for recognition of one speaker form another. Unlike speaker recognition using vocal tract

parameters, here we are able to recognize speakers of same vocal tract characteristics. Also, pitch contour will not be affected by spectral characteristics of recording and transmission systems. Hence it can be used more effectively in real life situations like recognition of speakers from telephone speech.

2.4 Outline of the present work

Our study aims at the investigation of the properties of intonation patterns in the Indian language, Hindi. Unlike many other languages, Indian languages exhibit phonetic property. That is, their ***graphemes*** (orthographic representation of the language) and ***phonemes*** (units of speech sound) exhibit a definite correlation. Even though the work proposed here is for Hindi, this approach can be extended to any other Indian language due to the similar phonetic behavior of Indian languages.

Intonation patterns in Hindi show some regular features. Like many other languages discussed above, F_0 contour of declarative sentences in Hindi decline gradually with time and for interrogative sentences in Hindi F_0 contour rises towards the end. These backdrop declination or rising is characterized by local falls and rises. F_0 contour gets modified across syntactic boundaries and is called resetting of F_0 contour. The resetting is used as a marker for phrase boundaries and is accompanied by a pause. The intonation pattern of the utterance is also affected by the phonetic factors of constituent units. Intonation knowledge for continuous speech accounts for all these factors.

Proper incorporation of intonation knowledge will enhance the intelligibility and the naturalness of synthetic speech significantly. In a text-to-speech system intonation refers to the periodicity of the source for voiced speech sounds. By exploiting the properties of F_0 contour, we are able to hypothesize some word boundaries and function words from continuous speech in Hindi. The properties of intonation patterns will also help us to differentiate one speaker from another from their speech characteristics. One advantage of using intonation knowledge for word boundary hypothesization and speaker recognition is its robustness towards noise.

The whole research study reported in this thesis can be described into two parts. The first part discusses the issues in the acquisition of intonation knowledge. We consider the issues related to different global, local and phonemic properties of intonation patterns in Hindi in the order of their importance in continuous speech. Incorporation of intonation knowledge in various speech systems, such as text-to-speech, speech-to-text and speaker recognition system is discussed in the second part. Chapters 3, 4 and 5 discuss the issues in the acquisition of intonation knowledge. Incorporation of intonation knowledge in various speech systems is discussed in chapters 6, 7 and 8.

2.5 Summary

This chapter reviewed the properties of intonation patterns in various languages in the context of speech systems. Intonation properties are classified into global properties, local fall-rise patterns, resetting of F_0 contour and segmental properties. Each of these properties were reviewed with respect to the available literature. The studies on the intonation properties in Indian languages are very limited. There is hardly any study which discuss the properties of intonation patterns in the context of speech systems for Hindi. Applications of intonation knowledge in speech systems, such as text-to-speech, speech-to-text and speaker recognition systems were also reviewed. Finally, the outline of the present study was discussed.

Chapter 3
INTONATION KNOWLEDGE FOR SIMPLE DECLARATIVE AND INTERROGATIVE SENTENCES IN HINDI

3.1 Introduction

Pitch contour for a natural utterance can be extracted by several methods. For the analysis, we have used two algorithms: one based on simplified inverse filter tracking (Markel, 1972) and the other on the properties of group delay functions (Yegnanarayana, Murthy & Ramachandran, 1991). Intonation pattern is defined as the variation of F_0, the acoustic parameter of pitch over time. Properties of intonation patterns change with respect to the type of the sentence. This chapter discusses methods used for the extraction of pitch contour and the properties of intonation patterns for simple declarative and interrogative sentences for Hindi.

F_0 contour of a declarative sentence decline gradually with time whereas for interrogative sentences F_0 contour rises towards the end. This declinatiodrising tendency is accompanied by local falls and rises which are determined by the phonological pattern of the constituent words. It is possible to capture the features of these declinatiodrising tendency and local fall-rise pattern in some systematic fashion.

This chapter is organized as follows: Section 3.2 discusses the methods used for the extraction of pitch contour. In Section 3.3, we discuss the properties of intonation patterns in Hindi. It includes a discussion on the global properties of Fo contour in simple declarative sentences and interrogative sentences in Hindi and a discussion on the properties of local fall-rise patterns in Hindi.

3.2 Extraction of pitch contour

Separating source information from the speech signal involves several issues. They are summarized as follows: (1) The excitation function is switched on and off between voiced and voiceless excitation, and sometimes both excitations act simultaneously. (2) The amplitude of the speech signal changes continuously. (3) The voiced excitation function can vary its carrier frequency, that is, the fundamental frequency (F_0) of the speech signal within a wide range. These features are deliberately used by the speaker to support the process of phonation with the vocal tract. They yield a complexity to the voice source signal which makes the estimation of such parameters difficult (Hess, 1983).

The voice source parameters include the type of phonation (voiced or voiceless) and the measure of periodicity (pitch period or Fo) of the speech wave if the signal is voiced. There exist several algorithms to determine the pitch period of a speech signal (e.g., Rabiner, Cheng, Rosenberg & McGonegal, 1976; Hess, 1983; Yegnanarayana, Murthy & Ramachandran, 1991). They are based on autocorrelation functions, cepstrum, simplified inverse filter tracking, parallel processing time domain methods, data reduction, linear predictive coding, average magnitude difference functions, group delay functions, etc.. Of these we are using the pitch extraction methods based on simplified inverse filter tracking and the properties of group delay functions. In the following sections we discuss each of these methods in detail.

3.2.1 Simplified Inverse Filter Tracking (SIFT) algorithm for pitch extraction

This method is based on an inverse filter formulation which retains the advantages of both the autocorrelation and the cepstral analysis techniques. Fig.3.1 is a demonstration of this technique. Fig.3.1a is the speech segment (duration 25.6 msec) for extracting pitch. The Fig.3.1b is the output of inverse filter analysis normalized to unity at the origin. Fig.3.1c shows the output normalized to the pitch peak after ignoring first 2 msec. This output sequence is defined as the autocorrelation of the inverse filter output and thus can be normalized on the ordinate corresponding to correlation values from -1.0 to

+1.0. There is a sharp peak corresponding to the correlation value of 0.8 at 7.8 msec. The pitch period is defined as the location of this peak and generally it exhibits the largest correlation over all samples except at the origin. Since it is always possible to normalize the output, and the data values are physical interpretation of the correlation, it should be possible to define a simple voiced-unvoiced decision based upon a fixed threshold value (Markel, 1972).

Fig.3.2 shows the block diagram for the pitch extraction based on SIFT algorithm. The speech waveform is first prefiltered by a low pass filter with a cut off frequency of 0.8 kHz. Sample the filter output at 2 kHz rate. Sampled speech is then segmented into analysis frames of suitable size. We used 256 samples with a shift of 64 samples between successive frames for extracting pitch. Perform a short-term autocorrelation analysis on the data and find out the first five terms (r_j, $j = 0, 1,..4$). The inverse filter coefficients a_i can be obtained by solving a set of linear equations

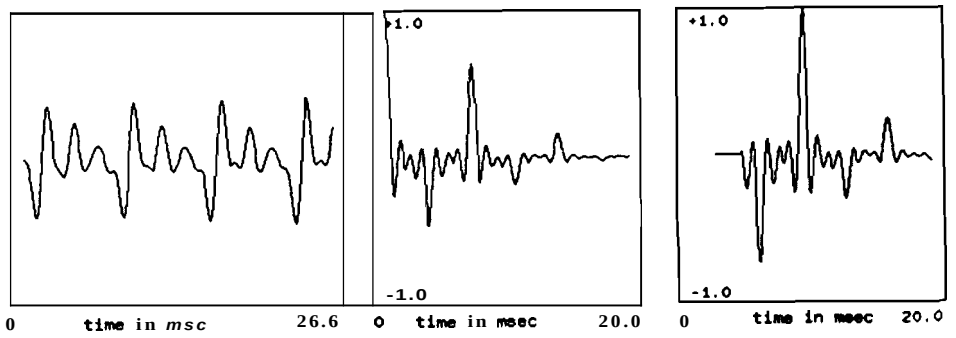
$$\sum_{i=1}^{4} a_i r_{i-j} = -r_j, \quad j = 1, 2, ...4.$$

The inverse filter is defined by

$$A(z) = 1 + \sum_{i=1}^{4} a_i z^{-1}$$

Thus by knowing a_i, the inverse filter output can be calculated. The output correlation sequence from which F_0 is estimated is then calculated as the autocorrelation sequence of inverse filter output. The largest peak of autocorrelation sequence within the specified limits (after ignoring first 2 msec) corresponds to the pitch period. Interpolation is applied to the region of the peak, and then voiced/unvoiced decision is made based upon the interpolated peak. If the segment is voiced, the reciprocal of the location of interpolated peak is defined as the F_0 of the segment.

SIFT algorithm is widely used for extraction of F_0 contour in speech analysis. Its advantages are: 1) the voiced/unvoiced decision algorithm is very simple; 2) implementation requires only elementary arithmetic operations; 3) algorithm is very efficient computationally and 4) it gives accurate value of Fo for



(a) (b) (c)

Fig.3.1. Segmental analysis in simplified inverse filter tracking (SIFT) algorithm.

(a) A segment of voiced speech (25.6 msec).

(b) Output of autocorrelation analysis normalized to unity art the origin.

(c) Output of autocorrelation analysis normalized to unity after zeroing the first 2 msec. Location of the maximum peak corresponds to the pitch period.

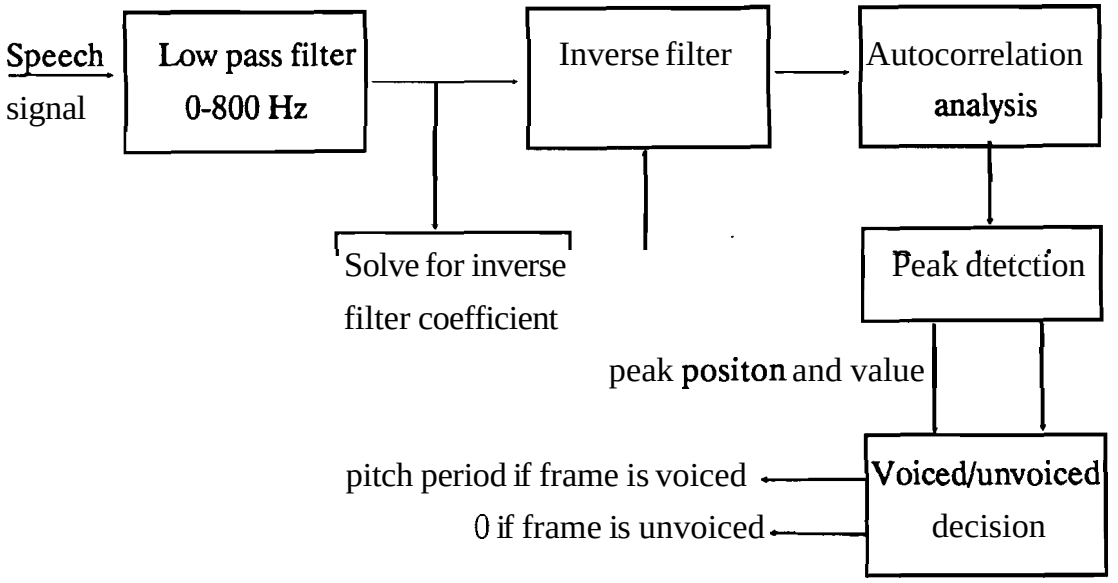


Fig.3.2. Block diagram for pitch estimation using simplified inverse filter tracking algorithm

clean speech. But this algorithm does not guarantee an error free analysis. Also the **voiced/unvoiced** decision is based on threshold logic which does not work accurately at all the time. Another disadvantage of this algorithm is the performance in noisy speech input conditions. If the input speech is noisy, then this algorithm fails miserably.

3.22 Pitch extraction based on the properties of group delay functions.

Most of the speech processing methods use spectral magnitude characteristics of short-term analysis segment of speech data. Computation of Fourier transform (FT) magnitude in some form or other dominated in all these methods, although signal representation is complete only if both FT magnitude and FT phase are used. FT phase is largely ignored in speech processing methods party because it is not well understood and partly because the difficulty in processing FT phase directly due to wrapping problems (values being in the range $\pm\pi$) in phase computation via discrete Fourier transform. The negative derivative of FT phase, or group delay functions can be directly computed from the signal. Therefore instead of processing of FT phase directly, it can be processed indirectly through group delay functions (Yegnanarayana & Ramachandran, 1992).

Given a segment of speech signal, $x(n)$, $n = 0, 1, \ldots N-1$, the group delay function is computed as follows:

Let $X(k)$ and $Y(k)$ be discrete Fourier transform of the sequences $x(n)$ and $nx(n)$, respectively. The samples of group delay function is given by

$$\tau(k) = \frac{(X_R(k)Y_R(k) + X_I(k)Y_I(k))}{(X_R^2(k) + X_I^2(k))}$$

where subscripts R and I refer to the real and imaginary parts, respectively (Oppenheim & Schaffer, 1975).

Let $|V(k)|^2$ be an estimate of zero spectrum. This is derived by flattering the magnitude spectrum either by linear prediction or by cepstral analysis (Rabiner & Schaffer, 1978). The modified group delay function is given by

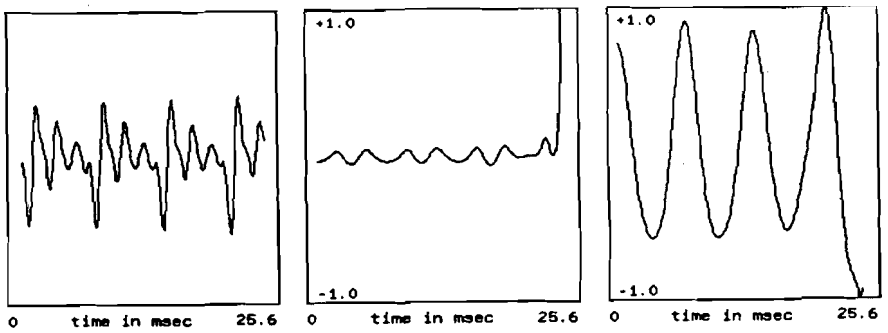
$$\hat{\tau}(k) = \tau X(k)|V(k)|^2$$

The periodic glottal pulse excitation in voiced segments manifests as sinusoids in the frequency domains. In other words, magnitude spectrum $|X(k)|^2$ of a voiced segment contains a sinusoidal component corresponding to pitch, besides peaks due to formants and random fluctuation due to excitation and additive noise. The problem of pitch extraction is simply the estimation of the frequency of the sinusoid in $|X(k)|^2$ even when there is distortion and noise. We can consider high SNR (signal to noise ratio) portions of $|X(k)|^2$ as signal and compute the modified group delay function for the signal. The peak in the modified group delay function corresponds to pitch period. Fig.3.3 shows the group delay processing of a short segment of voiced speech data. **Fig.3.3a** shows the speech segment of duration 25.6 msec. **Fig.3.3b** is the corresponding modified group delay function. **Fig.3.3c** shows the modified group delay functions for the zero spectrum of the voiced segment of **Fig.3.3a.** It is possible to determine the pitch period from the locations of the peaks in the **Fig.3.3c.**

The algorithm for extracting pitch contour using the properties of group delay functions is given in **Fig.3.4.** The algorithm considers the first four peaks in modified group delay functions and arrange their positions in ascending order. Compute the distance between the adjacent peaks. The maximum of these distances corresponds to the pitch period. This algorithm extract the pitch period correctly for most of the voiced segments and the pitch values are random for unvoiced and silence regions of speech. This method of pitch extraction works well for noisy speech. It is due to robustness of features of phase in noisy input conditions.

3.3 Properties of intonation patterns in Hindi

For the present analysis, we have used reading style of speech. While collecting the data, we instructed the speakers to read the sentences in neutral phonetic context to avoid the effects due to semantics. A corpus of **500** sentences was read out by two adult male native speakers of Hindi. Speech was digitized to 12 **bits/sample** at a sampling rate of 10 kHz. A 256 sample analysis frame with a shift of 64 samples was used for extracting pitch. In the following sections we



(a) (b) (c)

Fig.3.3. Segmental analysis in group delay processing of speech signals.

(a) A segment of voiced speech (25.6 msec).

(b) Group delay computed from the speech signal.

(c) Modified group delay computed from the group delay function. Let **Pi, P_2, P_3** and **P_4** are the peak positions.

For each frame perform the following steps:

1. Arrange peak positions **Pi, P_2, P_3** and **P_4** in ascending order.
2. Compute the distances **D_1 = Pi-0, D_2 = P_2-P_1, D3 = P_3-P_2** and **D4 = P_4-P_3.**
3. The maximum of **D_1, D_2, D3** and **D_4** corresponds to pitch period.

Smooth the pitch contour using a moving average of 15 points.

Fig.3.4. Algorithm for estimating pitch period using the properties of group delay functions.

discuss the global and local features of intonation patterns in Hindi. It covers the declination/rising tendency, fall-rise patterns and the various other observations in each of this attributes.

33.1 Declination/rising tendency

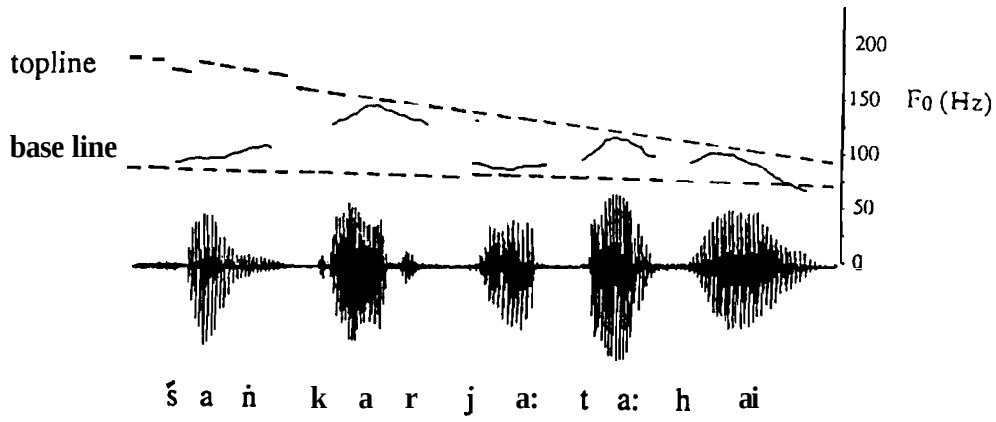
Like in many other languages, Fo contour for simple declarative sentences in Hindi has a tendency to decline gradually during the utterance. For interrogative sentences in Hindi, Fo contour rises towards the end. Properties of intonation patterns for declarative and interrogative sentences in Hindi are discussed in the following sections.

3.3.1.1 Intonation pattern for simple declarative sentences

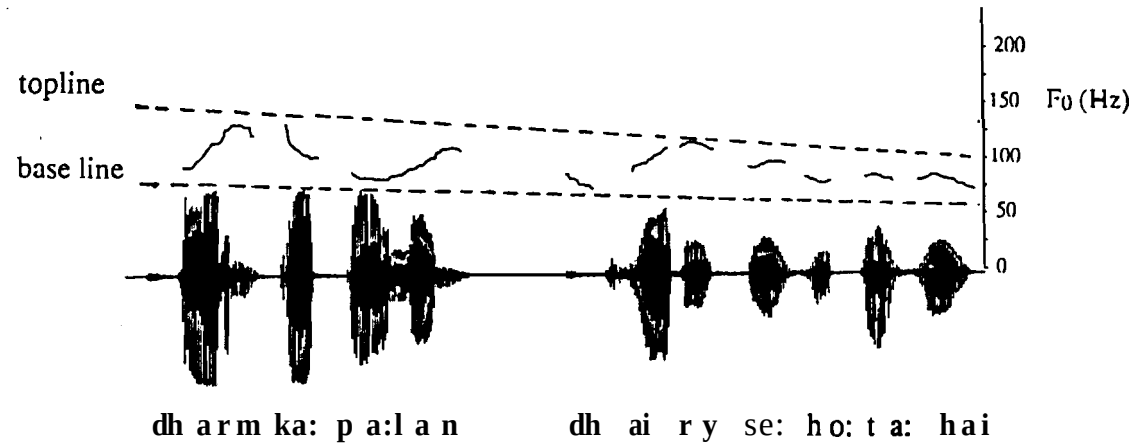
Declination of F_0 contour in a Hindi sentence is characterized by local falls and rises. These falls and rises fluctuate between two abstract lines -- a top line and a base line, drawn near or through all maxima and minima values of F_0 contour in a sentence, respectively. The repeated succession of falls and rises of F_0 contour are called valleys and peaks, respectively. The difference between a valley and next peak is called the range of F_0 contour. The range of F_0 contour decreases with time. That is, both top line and base line monotonically decrease and slope of the top line is steeper than the base line.

The text of an utterance can be divided into words. Words can be content *words* or *function* words. Content words are defined as semantically meaningful words. Words which have only grammatical value are called function words. In a neutral declarative sentence in Hindi, the maximum value of F_0 contour will be located in the stressed syllable of the first content word itself. In the connected speech the content word together with the preceding or the following function words, if any, form a pitch accent group called prosodic word under certain conditions. These will be determined by the rhythmic factors and other linguistic constraints.

Speech waveform and the corresponding F_0 contour for natural utterances of simple declarative sentences are shown in Figs.3.5. The F_0 contour starts at the initial syllable of the first word rises towards the next target, that is, the final



(a)



(b)

Fig.3.5. Speech waveform and F_0 contour for simple declarative sentences

(a) */šaṅkar ja:ta:* **hail** (Shankar goes)

(b) */dharm ka: pa:lan dhairy se: ho:ta:* **hail** (Patience is required to follow religion)

The F_0 contour declines towards the end of the utterance and is characterized by local falls and rises. These falls and rises fluctuate between two abstract lines - a top line and a base line.

syllable of the first content word. The F_0 rises and falls damp off towards the end of the utterance. It is possible to draw a line connecting all the peaks (top line) and another line connects all the valleys (base line). Both lines decline monotonically and converge towards the end.

3.3.1.2 Intonation patterns for interrogative sentences

Interrogative sentences in Hindi can be broadly classified into two. They are: 1) yes-no type questions and 2) question-word type questions. Yes-no type interrogative sentences in Hindi have the same grammatical structure as declarative sentences, except that optionally question may include the question word */kya:/* (what) usually at the beginning of the sentence. Thus declarative sentences may be distinguished from this type of interrogative sentences by prosody alone, if we leave aside the context of dialog. Question-word type interrogative sentences expect detailed answers and are marked in Hindi with any one of a set of interrogative words in Hindi. The commonly used interrogative words in Hindi and their corresponding English words are given in Table 3.1.

Question word	Meaning
1. *kya:*	what
2. *kab*	when
3. *kahã:*	where
4. *kiska:, kiske:, kiski:*	whose
5. *kitna:, kitne:, kitni:*	howmany
6. *kaisa:, kaise:, kaisi:*	how
7. *kaun*	who
8. *kaunsa:, kaunse:, kaunsi:*	whose
9. *kyõ:*	why

Table 3.1. Question words in Hindi

Intonation patterns for both types of interrogative sentences are different. Even though the global properties of F_0 contour are different for questions, the local attributes such as the fall-rise patterns remain same.

Questions expecting yes/no answers (yes-no type interrogative sentences) have a continuous rise in the F_0 contour. That is, both the top line and the base line have positive slope. Figs.3.6 shows the speech waveforms and the Fo contours for yes-no type questions. F_0 contour starts at the initial syllable of the first word and rise towards the next target, that is, the final syllable of the first content word. The F_0 falls and rises continues till the end of the utterance. But the magnitude of the valley and the peak increase with time. Hence the base line and the top line rise upward from left to right as shown in figure.

The intonation pattern for question-word type interrogative sentences exhibit a dual nature. The top line and the base line decline gradually up to the question-word and then rise towards the end. Figs.3.7 shows the speech waveforms and F_0 contours for question-word type interrogative sentences. In both the cases F_0 contour declines up to the question word (*/kaun/* (who) in Fig.3.7a and */kya:/* (what) in Fig.3.7b) and then rises toward the end. From the figure it can be seen that the local fall-rise patterns will not change with respect to type of the sentence.

3.3.1.3 Prediction of intermediate peaks and valleys

F_0 contour for an utterance is modeled with respect to the top line and the base line as discussed earlier. Since the aim was to capture the properties of F_0 contour by some empirical formula, a significant issue involved is the normalization for individual differences among phonetic segments, speakers and sentences. For instance, the inherent F_0 of vowels change with its phonetic properties as well as the properties of adjacent consonants (Detailed discussion on these issues are given in Chapter 5). Changes in F_0 contour due to these properties will range up to 30 Hz.

In order to predict the intermediate peaks and valleys for simple declarative sentences, we have selected 10 simple declarative sentences uttered by two adult

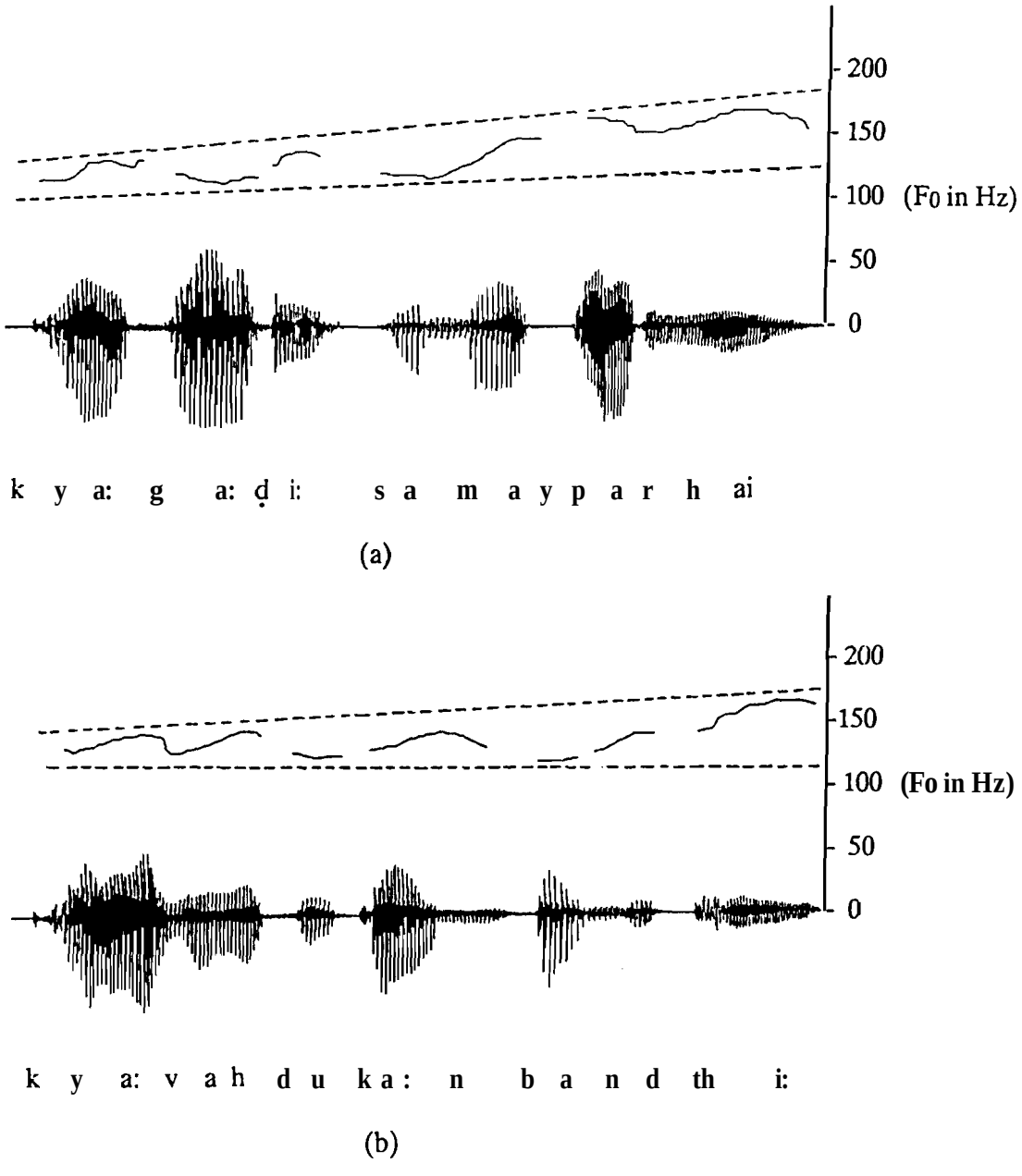
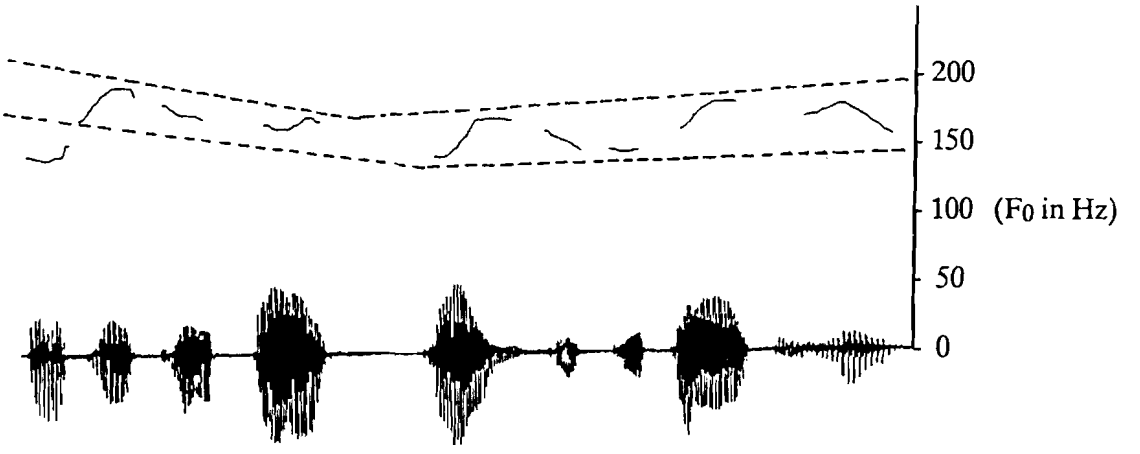


Fig.3.6. Speech waveform and F_0 contour for yes-no type interrogative sentences

(a) ***lkya: ga:ḍi: samay par hail*** **(Whether train is on time)**

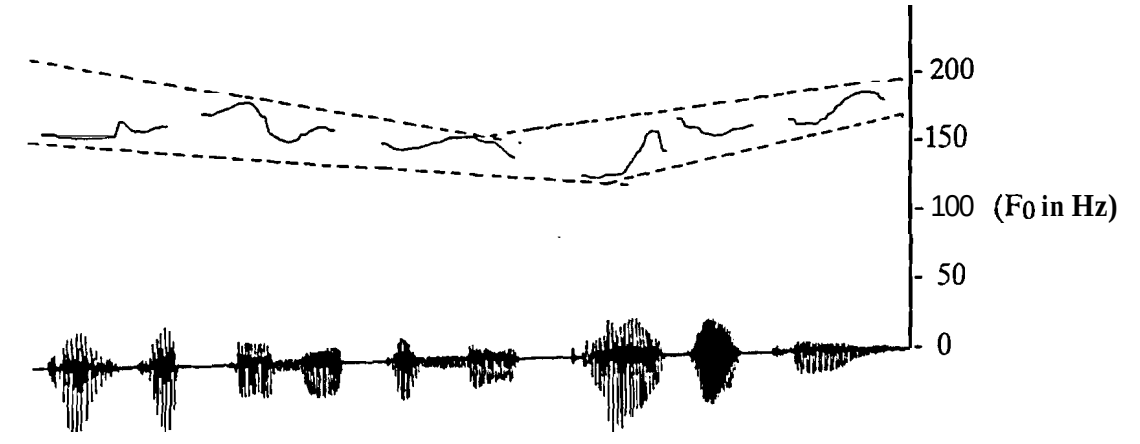
(b) ***lkya: vah duka:n band thi:/*** **(Whether the shop was closed)**

The falls and rises of Fo contour rise towards the end of the utterance.



a j ay ke: p a: s kau n s i: k i t a: b h ai

(a)



k a l k a t t e: m ẽ: t u m n e: k y a: d e: kh a:

(b)

Fig.3.7. **Speech waveform and F0 contour for question-word type interrogative sentences**

(a) */ajay ke: pa:s kaun si: kita:b hail* **(Which book is with Ajay)**

(b) */kalkatte: mẽ: tumne: kya: dekha:/* **(What have you seen in Calcutta)**

The top line and the base line decline gradually up to the question word and then rise towards the end.

male native speakers of Hindi. The set of sentences are given in Appendix B1. In all the sentences the number of prosodic words are same. The maximum and the minimum values of Fo of each content word fluctuate between the top line and the base line. The values of F_0 of valleys and peaks of the first and the final words are selected as reference values. If Po is the initial peak and Pn is the final peak at timings To and T_n, respectively, then an intermediate peak at time Ti can be expressed as

$$Pi = Po + (Ti - T0) * \frac{(P0 - Pn)}{(T0 - Tn)}$$

The tabular results of this experiment for valleys and peaks of Fo contour are given in Appendices C1 to C4. In all the cases the error is less than 10% of the actual intermediate values, keeping the range behavior same. The results of our analysis show that the model of F_0 contour with two reference lines has faithfully captured the properties of declination of F_0 contour. Hence we can approximately predict the intermediate peaks and valley if we know the initial and the final values.

A similar experiment was conducted for yes-no type interrogative sentences. Here the top line and the base line are increasing with respect to time. But the rate of increasing (slope) is different for both the top line and the base line. The intermediate peaks and valleys can be modeled as in the case of declarative sentences. For this analysis we collected the utterances of 10 yes-no type interrogative sentences from two adult male native speakers of Hindi. All sentences have the same number of prosodic words (sentences are given in Appendix B2). The results for the prediction of intermediate peaks and valleys of yes/no type interrogative sentences are given in Appendices C5 to C8. Here also the model predicted the valleys and peaks of F_0 contour accurately.

3.3.1.4 Range of F_0 contour for prosodic words in sentences

The difference in F_0 between a valley and the following peak is called the range of Fo. Style, emphasis and position of a word in the sentence are some of the factors which can affect the range of Fo. In the following paragraphs, we

discuss the change in the range of F_0 for prosodic words in simple declarative and yes-no type interrogative sentences.

In simple declarative sentences both the top line and the base line have negative slope and converges towards the end. As a result the range of Fo diminishes with respect to time. In general, peaks of F_0 contour tend to decrease more rapidly than valleys. Experiments on music also suggested that an upward pitch change takes more time than a downward pitch change (Ohala & Ewan, 1973). It shows that it is easier to lower the pitch than to raise it. As a consequence of this tendency the increase in the amplitude of successive peaks of Fo does not entirely compensate for the decrease in valleys, resulting in the diminishing of range of F_0 contour.

The highest range of F_0 contour for simple declarative sentences generally occur for the first content word since it contains the highest peak on its stressed syllable. Similarly the lowest range occurs in the final word. We conducted an experiment to determine the range of F_0 for each prosodic word in 10 simple declarative sentences for two adult male-native speakers of Hindi. The results of our analysis are given in Appendices C9 and C10. All sentences have the same number of prosodic words (sentences are given in Appendix B1). The results also show the mean and standard deviation (SD). From the analysis it is obvious that the range decreases with respect to the position of the prosodic word for simple declarative sentences in Hindi.

The experiments for yes-no type interrogative sentences give a different picture. Here the speakers deliberately increase the tension of laryngeal muscles to convey the linguistic information. The top line and the base line increases with time. In contrast with the observations for declarative sentences, the final prosodic word has the maximum range and the first word has the minimum. These observations were checked with several yes-no type interrogative sentences for two speakers. Results of these observations for 10 such sentences are given in Appendices C11 and C12. When sentences end with monosyllabic words (e.g., /*hai*/ (is)), the range would be less due to the tapering effect and hence the standard deviation (11.15 for speaker 1 and 14.30 for speaker 2) is high

for the range of the final word. From the results it is obvious that the range of Fo increases with respect to the position of prosodic word for yes-no type interrogative sentences in Hindi.

33.2 Local fall-rise patterns

Peaks and valleys of F_0 contour for an utterance form local fall-rise patterns. The occurrence of falls and rises in F_0 contour can be explained as the tension of vocal folds and relaxation of the tension, respectively. This can be treated as the manner in which lexical entities are acoustically combined. Fall-rise patterns are super imposed on each prosodic word where as declination and rising manifest through out the utterance. They are operationally distinguishable by virtue of their different domain.

By analyzing large amount of data, we have observed some general features of local falls and rises of F_0 contour which are determined by the phonological pattern of the constituent words. In the following section we discuss the features of F_0 contour of prosodic words in Hindi.

3.3.2.1 Pitch accent patterns in Hindi

Fo contour of prosodic words in Hindi exhibits a regular pattern of a valley precedes each peak. Valleys and peaks correspond to the prominence of a particular syllable in a content word. A syllable in Hindi has a vowel obligatorily and maximally three consonant on its left (onset) and three consonant on the right (coda) optionally (Ohala, 1983). Peaks and valleys of Fo contour can appear at any point within the region of vowel or any of the voiced constituents of the syllable. The assignment of valleys and peaks for prosodic words in Hindi can be described as follows.

3.3.2.1.1 Monosyllabic words

Monosyllabic words in Hindi can be classified into content words and function words. For a monosyllabic content word the valley and the peak occur within the same syllable. The sequence of valley followed by a peak is maintained. Here the Fo contour raise steadily.

Function words are a class of words which bears only grammatical information. They are different from content words which are semantically important. The class of function words in Hindi includes case markers, post positions, complimentizers, negative markers, conjunctions, relative pronouns, etc. The number of function words are very few but they occur frequently. Most of the function words are monosyllabic though a few are disyllabic. The monosyllabic function words in our database are listed in Table 3.2.

Category	Function words
(a) Case markers and post positions	*-ne:, ka:, ke:, se, ko:, mẽ:,* me, par
(b) Complimentizer	*ki*
(c) Negative marker	*na*
(d) Conjunctions	*aur, ya:*
(e) Relative pronoun	jo:
(f) Emphatic markers	hi:

Table 3.2. List of Monosyllabic function words in Hindi

There are three types of valley-peak assignment for monosyllabic function words. Function words like relative pronouns and conjunctions have independent existence. Their pitch accent is same as that of monosyllabic content words. All other function words may conjoin with the preceding or the following noun phrase. Among these, some function words (e.g., post position */ko:/*) may get accented and therefore prosodically they conjoin with the previous content word. In such cases peak of the previous content word is shifted to the top of the function word. In the third category, function words are conjoined with content words in an unaccented position. F_0 contour for such function words decrease monotonously.

Fig.3.8 shows the fall-rise pattern for monosyllabic words. Fall-rise pattern for a monosyllabic content word */ham/* (we) is given in Fig.3.8a. Fig.3.8b shows

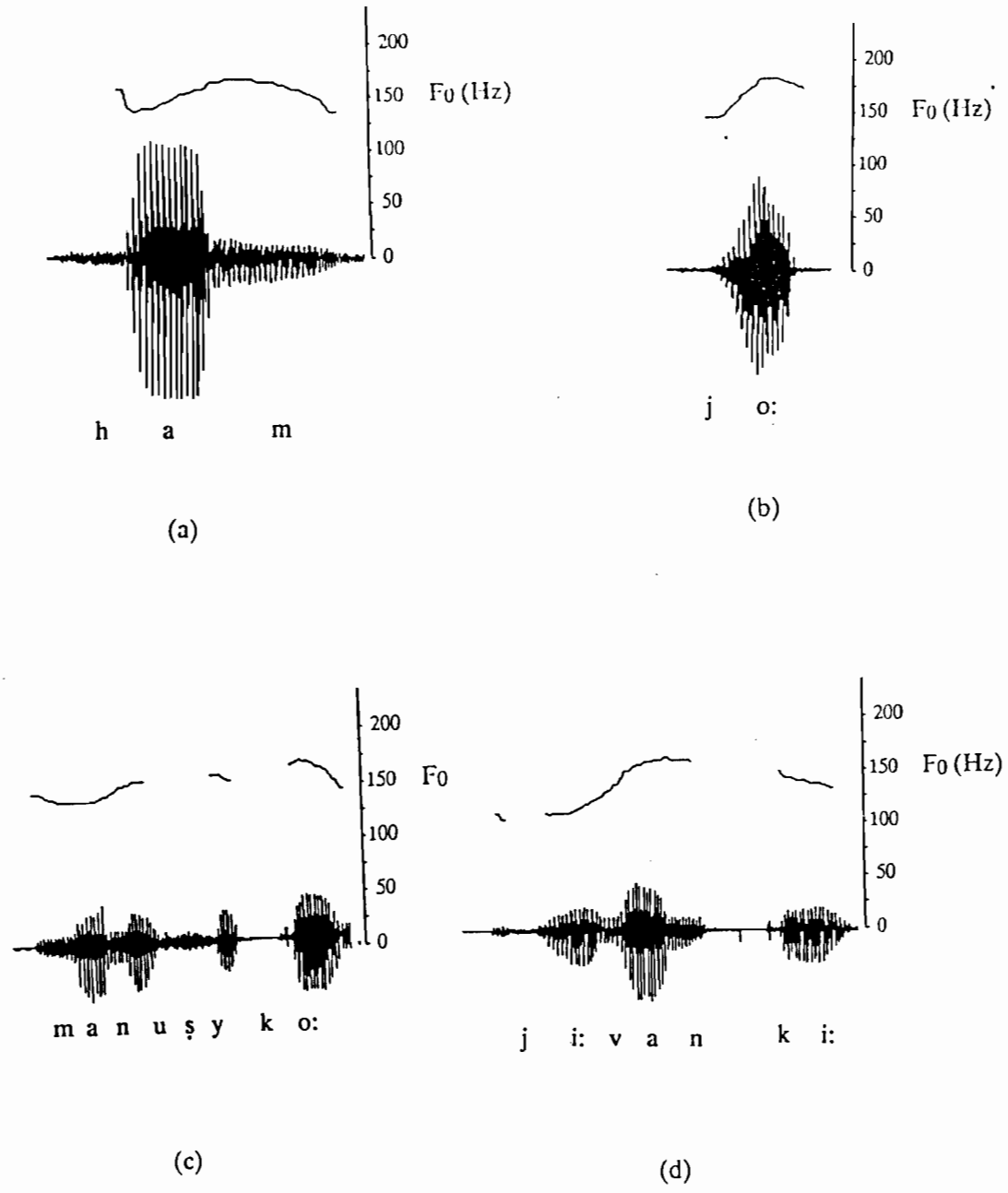


Fig.3.8. Pitch accent patterns for monosyllabic words

(a) Monosyllabic content word */ham/* (we)

(b) Monosyllabic function word */jo:/* (that)

(c) Monosyllabic function word */ko:/* accented with the previous content word */manuṣy/* (man)

(d) Monosyllabic function word */ki:/* conjoined with the previous content word */j:van/* (life)

In (a) and (b), the valley and the peak of F_0 contour occur in the same syllable. In (c), the peak of F_0 contour of the previous content word is shifted to the function word. In (d), F_0 contour of the function word decreases monotonously.

the fall-rise pattern for a monosyllabic function word */jo:/* (that) which shows independent existence. In Figs.3.8a and b, valley and peak occur on the same syllable. Fall-rise pattern of monosyllabic function word */ko:/* accented with the preceding noun phrase */manusy/* (man) is given in Fig.3.8c. The peak of the F_0 contour get shifted to the function word. Fig.3.8d shows the fall-rise pattern of a monosyllabic function word */ki:/* conjoined with the previous content word */ji:van/* (life) in an unaccented position. Here the fall-rise pattern for the content word does not change due to the addition of the function word.

3.3.2.1.2 Disyllabic and trisyllabic words

More than 70% of words in Hindi belong to this category. In disyllabic and trisyllabic words peak occurs on the final syllable and valley occurs on the initial syllable. However the exact target position is determined by several other factors. For instance, peak of the nucleus gets shifted to the coda (the consonant that follows the vowel nucleus) if the consonant is either lateral or nasal. This is true for all types of content words.

Fig.3.9 shows the fall-rise patterns for disyllabic words. Fall-rise pattern for the disyllabic word */pra:rtna:/* (prayer) is given in Fig.3.9a. Here valley occurs on the initial syllable */pra:r-/* and peak occurs on the final syllable */-tna:/*. Local fall-rise pattern for the disyllabic word */ke:val/* (absolute) is given in Fig.3.9b. The position of the peak get shifted to the coda since the lateral consonant */l/* is in the coda position. Fig.3.9c shows the local fall-rise pattern for the disyllabic word */ja:grit/* (awake). Here the coda is an voiceless consonant and hence no change in peak position.

Fig.3.10 shows the fall-rise pattern for trisyllabic words. Fall-rise pattern for the trisyllabic word */tu:pha:ni:/* (storm) is given in Fig.3.10a. The valley occurs on the initial syllable */tu:-/* and peak on the final syllable */-ni:/*. Fig.3.10b shows the fall-rise pattern for the trisyllabic word */pariṇa:m/* (consequence). The peak get shifted to the coda */-m/* since coda is nasal. Fig.3.10c shows the local fall-rise pattern for the trisyllabic word */gali: ke:/* formed by the combination of a disyllabic word */gali:/* (street) with an accented function word */ke:/*. The valley

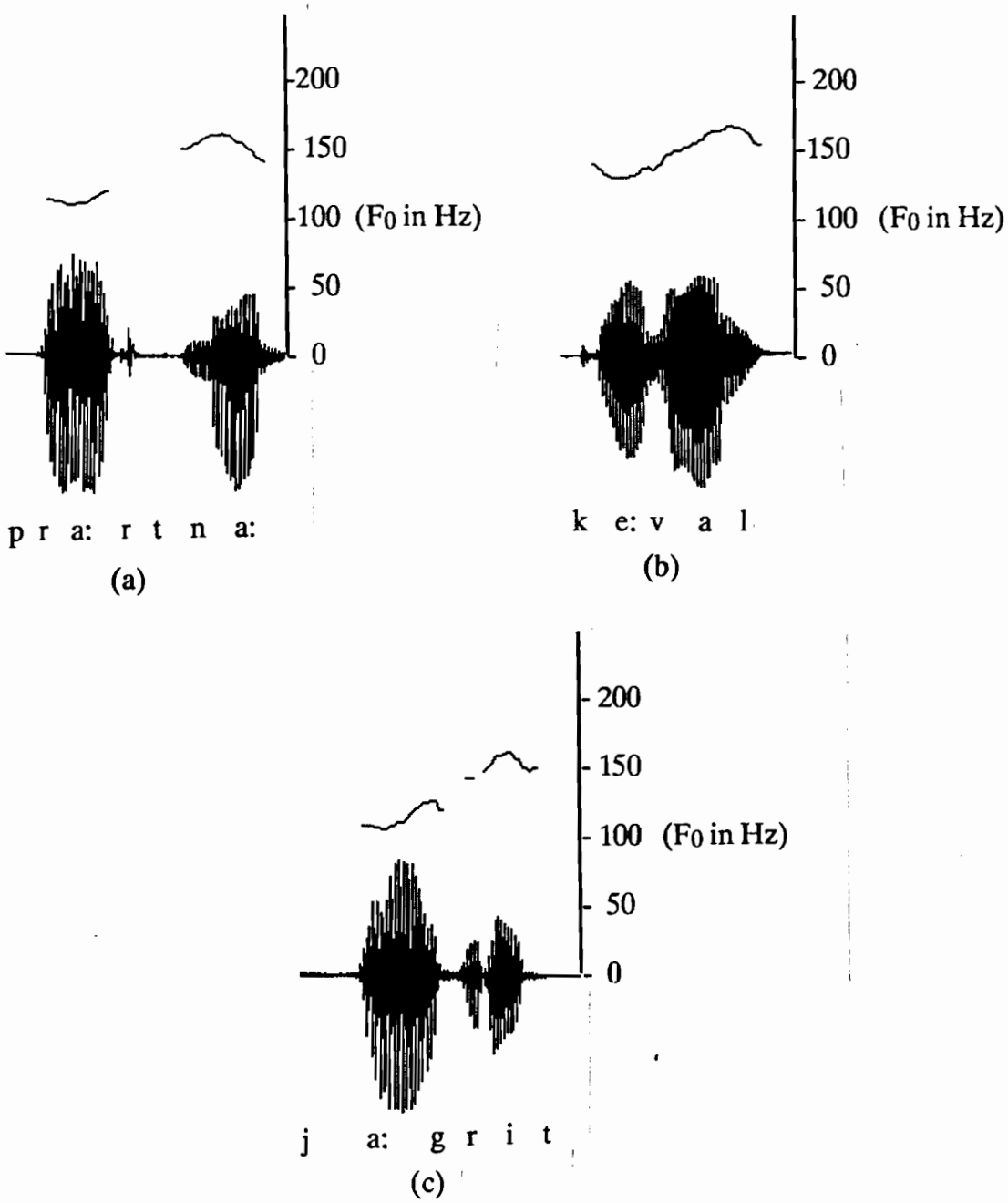


Fig.3.9. Pitch accent patterns for disyllabic words

(a) */pra:rtna:/* (prayer)

(b) */ke:val/* (absolute)

(c) */ja:grit/* (awake)

In all the cases valley occurs on the initial syllable and the peak occurs on the final syllable. In (b), the peak of F_0 contour is shifted to the coda since the coda is lateral.

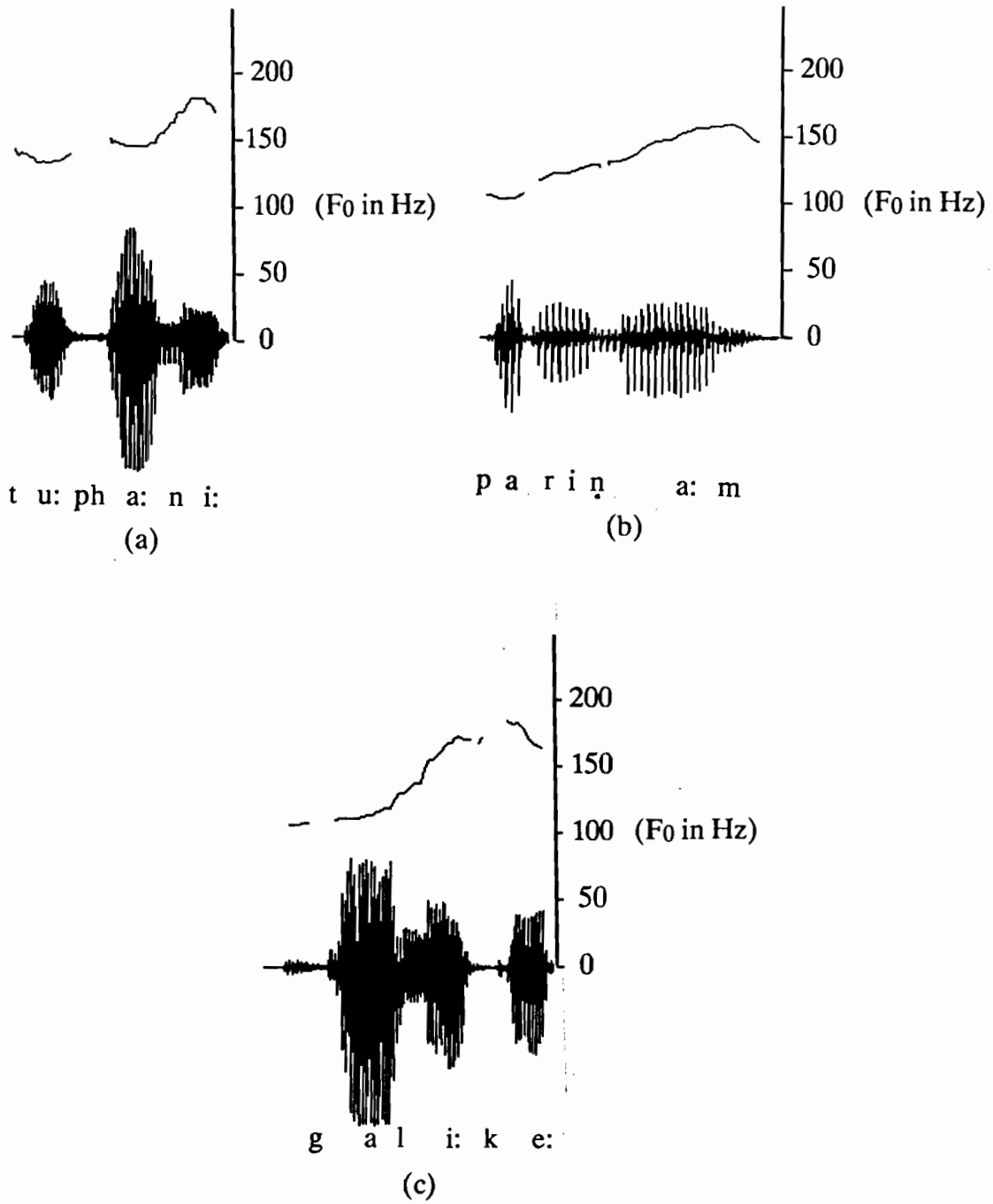


Fig.3.10. Pitch accent patterns for trisyllabic words

(a) */tu:pha:ni:/* (storm)

(b) */pariṇa:m/* (consequence)

(c) */gali: ke:/* (street)

In all the cases valley occurs on the initial syllable and the peak occurs on the final syllable. In (b), the peak of F_0 contour is shifted to the coda since the coda is nasal.

occurs on the initial syllable */ga-/* and the peak occurs on the function word */ke:/*.

3.3.2.1.3 Tetrasyllabic and pentasyllabic *words.*

Tetrasyllabic words show two types of patterns: (1) a valley on the initial syllable and a peak on the final syllable, and (2) the valley and peak occur on alternate syllables and hence characterized by two valleys and two peaks. The difference between two patterns are caused by the number of morphemes in the word. Pattern 1 is preferred when the word is monomorphemic (root), and pattern 2 is preferred when the word is bimorphemic. The fall-rise pattern for pentasyllabic word is similar to that of the pattern for a combination of disyllabic and trisyllabic words. Monomorphemic pentasyllbic words were not found in our database. Hence it is not possible to make any valid generalization. However, the occurrence of pentasyllabic and higher order words are very rare in Hindi. Table 3.3 shows some examples of polymorphemic words in Hindi and corresponding morphemes.

Fig.3.11 shows the fall-rise patterns for tetrasyllabic words. **Fig.3.11a** is a monomorphemic tetrasyllabic word */sahiṣṇuta:/* (tolerance) in which the valley occurs on the first syllable */sa-/* and the peak occurs on the final syllable */-ta:/*. Local fall-rise pattern for a bimorphemic tetrasyllabic word */antarya:mi:/* (omniscient) is given in **Fig.3.11b.** The morphemes are */antar-/* and */-ya:mi:/*. The valleys occur on the first and the third syllables (*/an-/* and */-ya:-/*) and the peaks occur on second and final syllable (*/-tar-/* and /-mi:/).

Fig.3.12 shows the fall-rise pattern for pentasyllabic words. **Fig.3.12a** shows the fall-rise pattern for a polymorphemic pentasyllabic word */smajhne:va:la:/* (the person who understands). The morphemes are */samajhne:-/* and */-va:la:/*, and the valleys occur on the first and the fourth syllables (*/sa-/* and */-va:-/*) and the peaks occur on the third and the fifth syllable (*/-jhne:-/* and */-la:/*). Fall-rise pattern for another bimorphemic pentasyllabic word */sarvaśaktima:n/* (Almighty) is given in **Fig.3.12b.** Here the morphemes are */sarva-/* and */-śaktima:n/*, and the valleys occur on the first and the third syllables (*/sa-/* and */- śa-/*) and the peaks occur on the third and the fifth syllable (*/-va-/* and */-man/*).

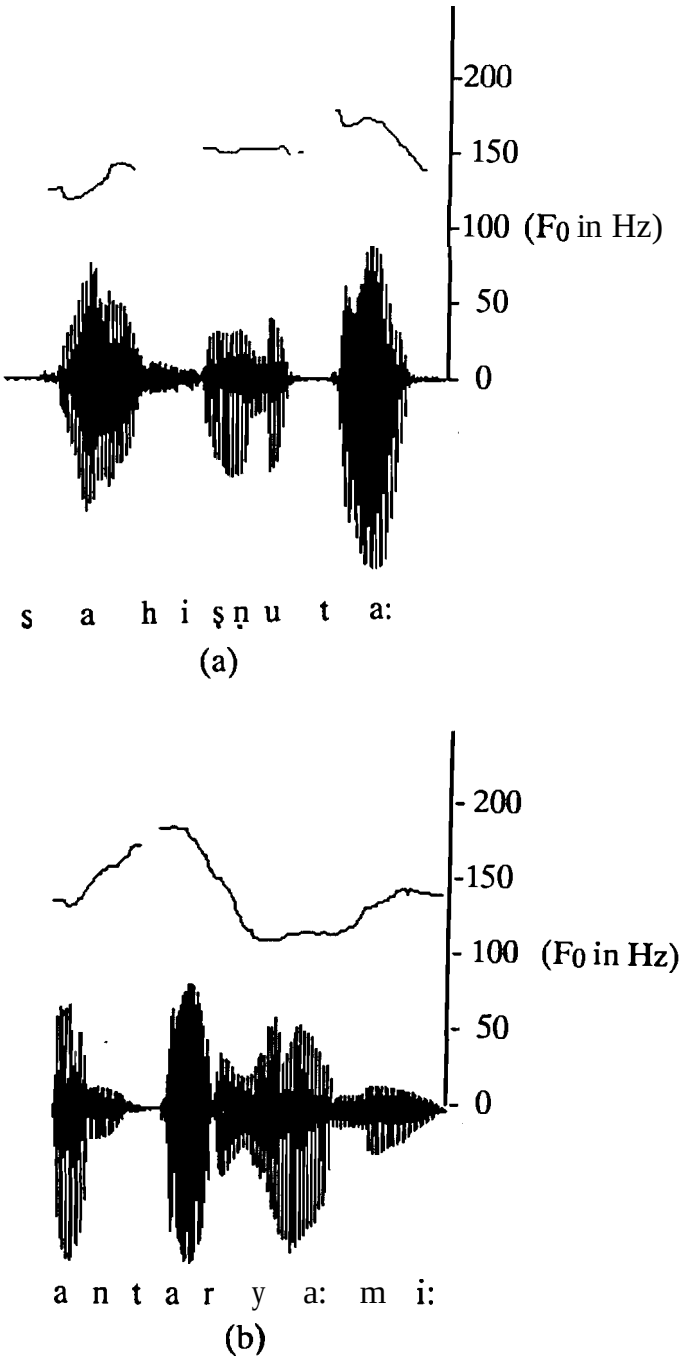


Fig.3.11. Pitch accent patterns for tetrasyllabic words

(a) */sahiṣṇuta:/* (tolerance)

(b) */antarya:mi:/* (omniscient)

The word (a) is monomorphemic and correspondingly the valley occurs on the initial syllable and the peak occurs on the final syllable. The word (b) is bimorphernic and hence the F_0 contour is characterized by two valleys and two peaks in alternate syllables.

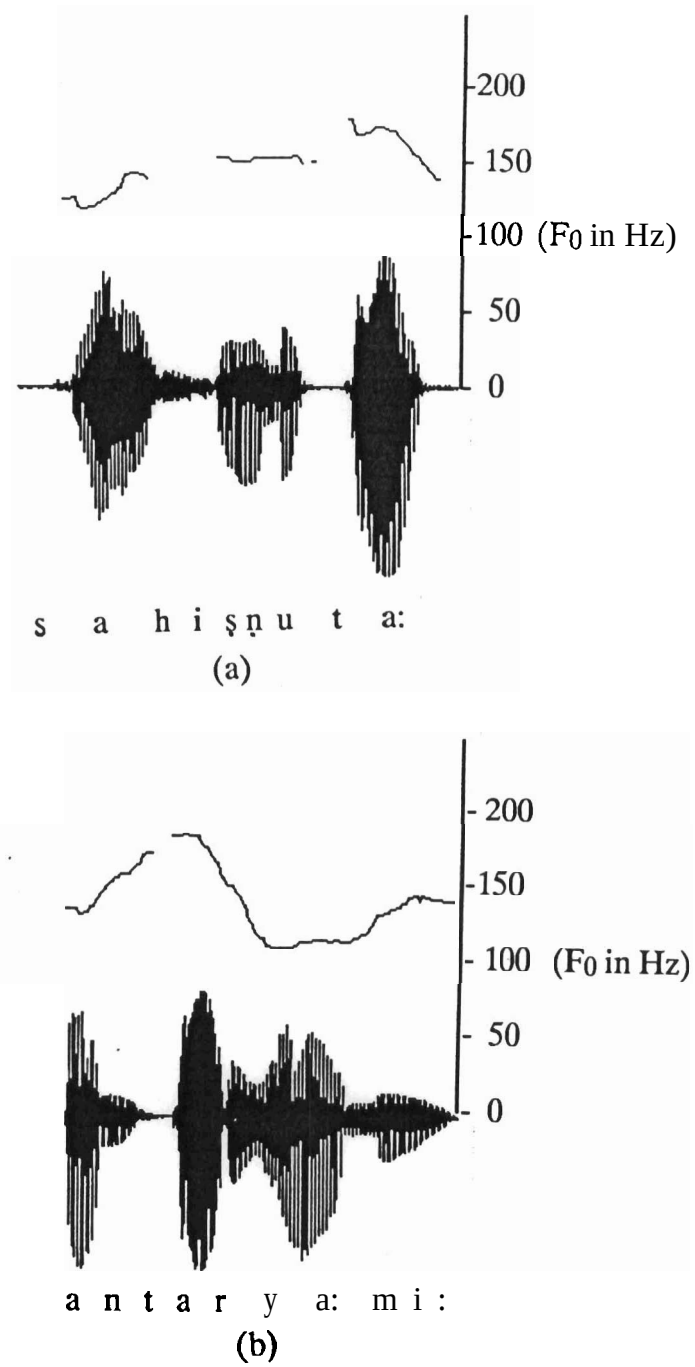


Fig.3.11. Pitch accent patterns for tetrasyllabic words

(a) */sahiṣṇuta:/* (tolerance)

(b) */antarya:mi:/* (omniscient)

The word (a) is monomorphemic and correspondingly the valley occurs on the initial syllable and the peak occurs on the final syllable. The word (b) is bimorphemic and hence the F_0 contour is characterized by two valleys and two peaks in alternate syllables.

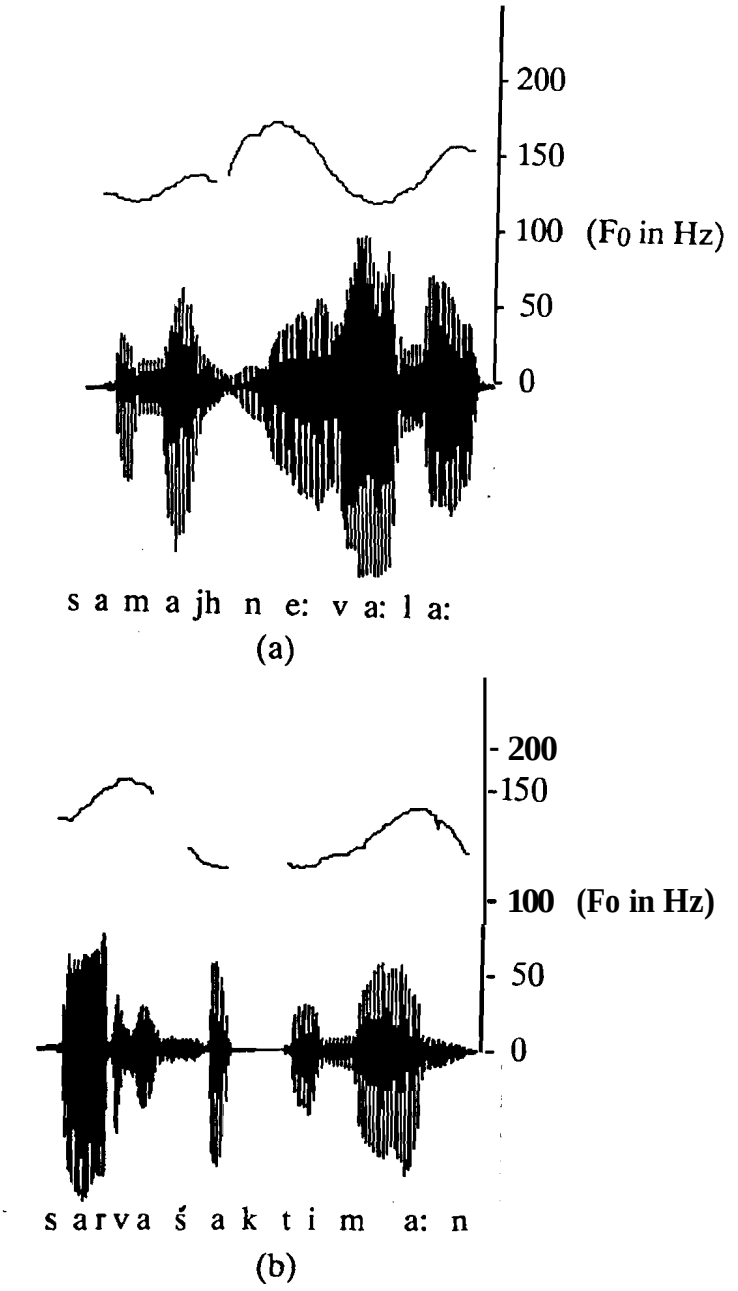


Fig.3.12. **Pitch accent patterns for pentasyllabic words**

(a) */samajhne:va:la:/* **(one who understands)**

(b) */sarvaśaktima:n/* **(alrnighty)**

Both the words are polymorphemic and correspondingly two valleys and two peaks occured in the F_0 **contour.**

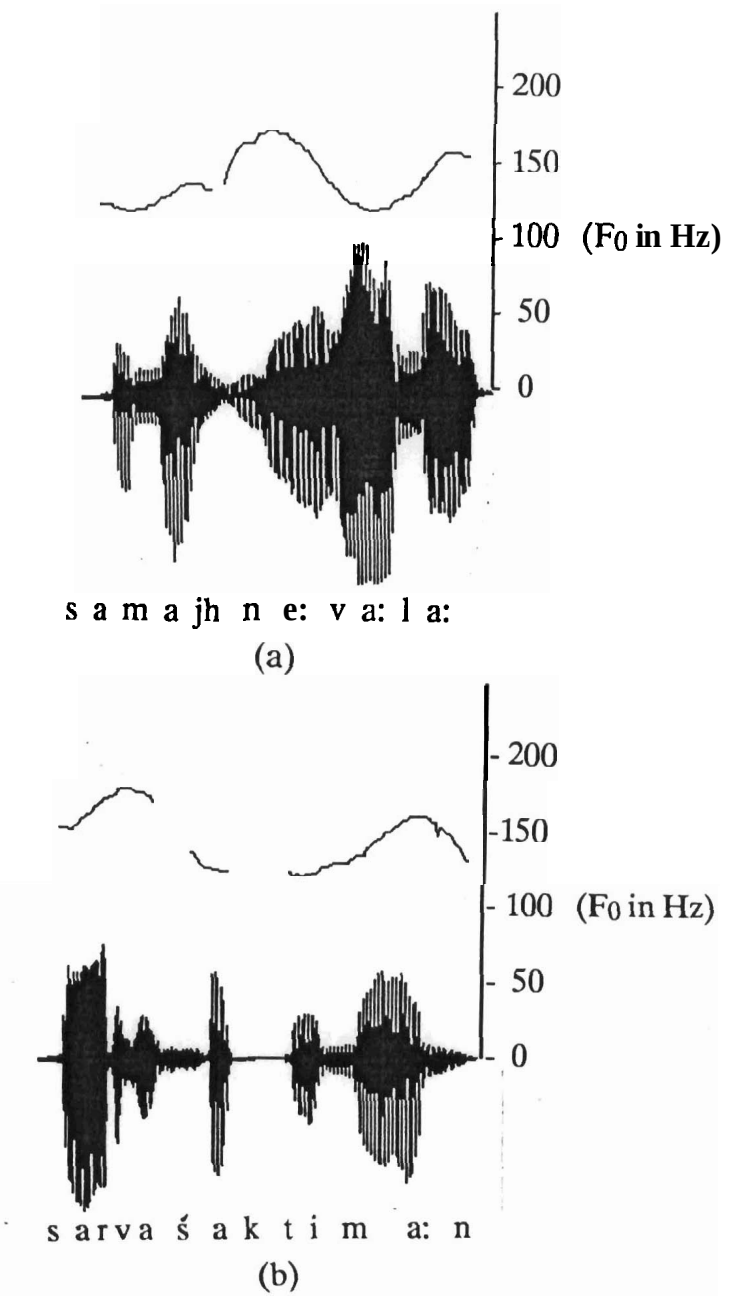


Fig.3.12.Pitch accent patterns for pentasyllabic words

(a) */samajhne:va:la:/* **(one who understands)**

(b) */sarvaśaktima:n/* (almighty)

Both the words are polymorphemic **and correspondingly two valleys and two peaks occured in the** F_0 **contour.**

Word	Morphemes
1. a:tmasamarpaṇ	a:tma + samarpaṇ
2. sarvaśaktima:n	sarva + śaktima:n
3. sa:rvajani:k	sa:rva + janik
4. e:ka:dhika:r	e:ka + adhika:r
5. kala:kauśal	kala: + kauśal
6. krama:nuku:l	krama + anuku:l
7. caturdaśi:	catur + daśi:
8. tapo:bhu:mi:	tapo: + bhu:mi
9. tilakmudra:	tilak + mudra:
10. daṇḍasamhita:	daṇḍa + samhita:
11. di:rghadarśita:	di:rgha + darśita:
12. na:takaśa:la:	na:taka + śa:la:
13. paripa:lak	pari + pa:lak
14. punarnirma:ṇ	punar + nirma:ṇ
15. bahuvacan	bahu + vacan

Table 3.3. Examples of poly morphemic words in Hindi

3.3.2.2 Effect of word order in Hindi

Hindi is a free word order language. However, the pitch accent patterns of words do not undergo any accent shift when the order of words is changed. For example, the sentence */usta:d ne: mina: ko: sita:r sikha:ya:/* (Master taught sitar to Mina) can be written in at least six different ways by changing the order of the subject (*/usta:d ne/*), direct object (*/mi:na: ko/*), indirect object (*/sita:r/*) and the verb phrase (*/sikha:ya:/*) in the sentence. The fall-rise pattern of the words remains constant irrespective of the change in the word order. However, the range between the valley and the peak vary, and this is determined by the

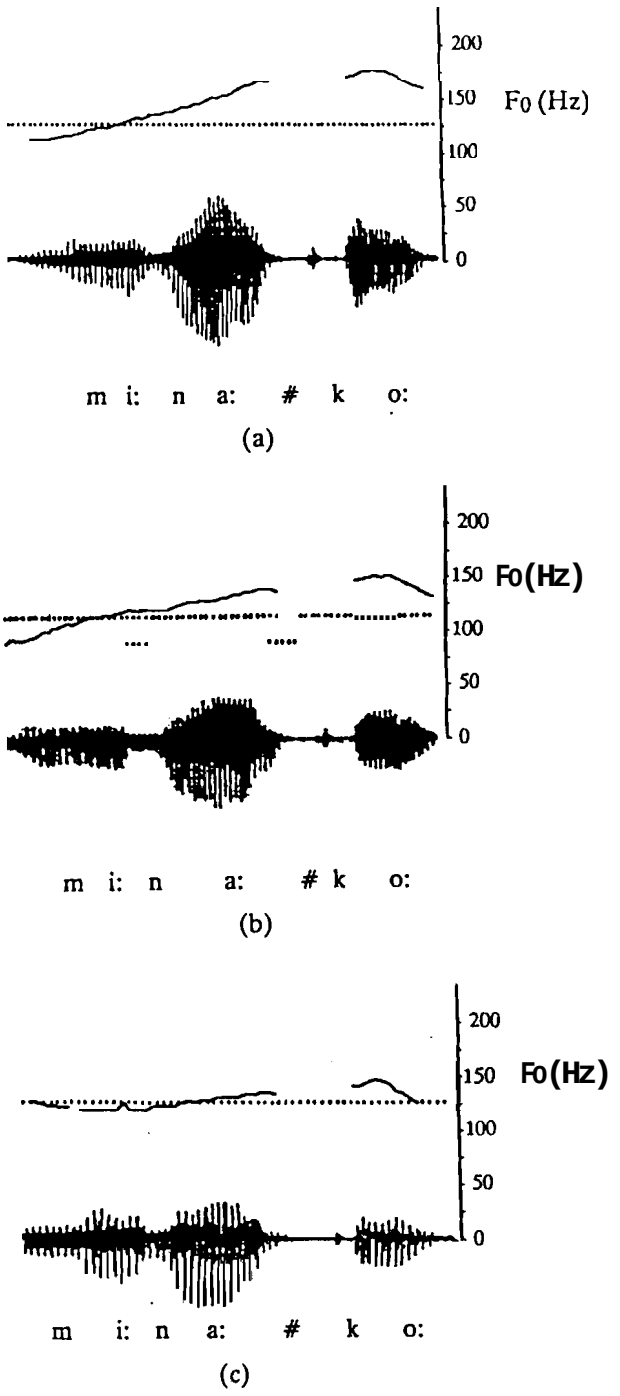


Fig.3.13. Pitch accent patterns for the prosodic word (*/mi:na: ko:/*) when it occurs in the following three positions in a sentence.

(a) */mi:na: ko: sita:r usta:d ne: sikha:ya:/*

(b) */sita:r mi:na: ko: usta:d ne: sikha:ya:/*

(c) */usta:d ne: mi:na: ko: sita:r sikha:ya:/*

(Master taught sitar to Mina)

The valleys and the peaks of Fo contour of a prosodic word does not change even if the word occurs at different positions in a sentence. But the range of F_0 contour changes according to the position.

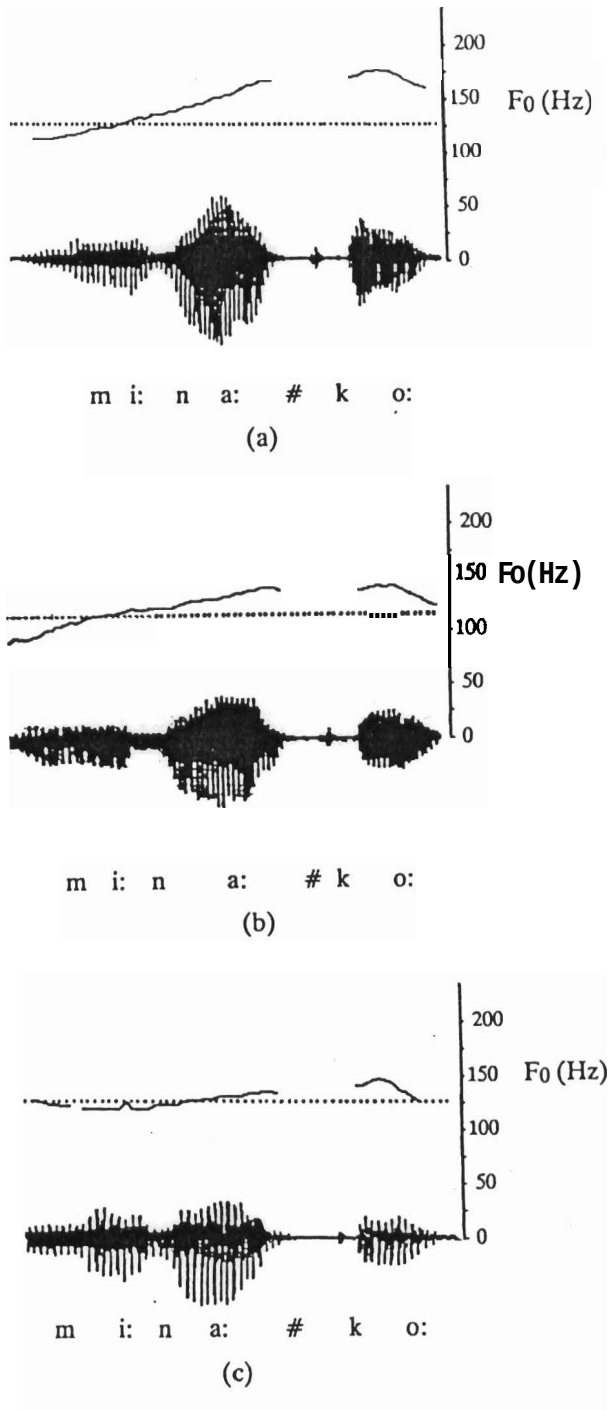


Fig.3.13. Pitch accent patterns for the prosodic word (*/mi:na: ko:/*) when it occurs in the following three positions in a sentence.

(a) */mi:na: ko:* *sitar* ***usta:d*** *ne: sikha:ya:/*

(b) */sita:r mi:na: ko: usta:d ne: sikha:ya:/*

(c) */usta:d ne: mi:na: ko: sitar sikha:ya:/*

(Master taught sitar to Mina)

The valleys and the peaks of F_0 contour of a prosodic word does not change even if the word occurs at different positions in a sentence. But the range of F_0 contour changes according to the position.

position of a word in the sentence. When a word occurs in the initial position of a declarative sentence, the range between the valley and the peak is around 30 to 40 Hz., whereas when the same word occurs in the final position the range is not more than 10 Hz. Fig.3.13 shows the constant fall-rise pattern of the same word (/*mina: ko:*/) with differing pitch ranges when the word occurs in different positions in a sentence.

3.4 Summary

This chapter discussed two methods of pitch extraction used in our analysis, one is based on simplified inverse filter tracking algorithm and the other is based on the properties of group delay functions. Properties of intonation patterns for simple declarative and interrogative sentences in Hindi were discussed. Fo contour for a declarative sentence declines gradually with time. Intonation patterns for interrogative sentences are of two types: F_0 contour rises continuously for an yes-no type interrogative sentence whereas for a question word type interrogative sentence, F_0 contour decreases upto the question word and then rises towards the end. This backdrop declination/rising tendency is accompanied by local falls and rises which are determined by phonological pattern of constituent words. Pitch accent patterns in Hindi show some regular features and they do not change with respect to the position of a word in the sentence.

Global properties of F_0 contour get modified across major syntactic boundaries. In the next chapter, we discuss the properties of intonation patterns for complex declarative and compound sentences in Hindi.

Chapter 4
INTONATION KNOWLEDGE FOR COMPLEX DECLARATIVE AND COMPOUND SENTENCES IN HINDI

4.1 Introduction

Properties of intonation patterns will change if the sentence consists of more than one syntactic clause. Such sentences can be called as complex sentences. In this chapter we discuss the intonation knowledge for complex sentences. Complex sentences in Hindi can be classified into complex declarative and compound sentences which are determined by subordinate and coordinate conjunctions in Hindi, respectively.

F_0 pattern gets modified across major syntactic boundaries. This is called resetting of F_0 contour. The resetting is used as a marker for phrase boundaries most of the time and it is accompanied by a pause. The part of utterance delimited by such a pause is called *intonational phrase*. Resetting of F_0 contour takes place both in valleys and peaks. But the magnitude of resetting differs in both the cases, and is determined by various constraints related to physiology, syntax and semantics. The amount of pause between syntactic clauses are also determined by these constraints.

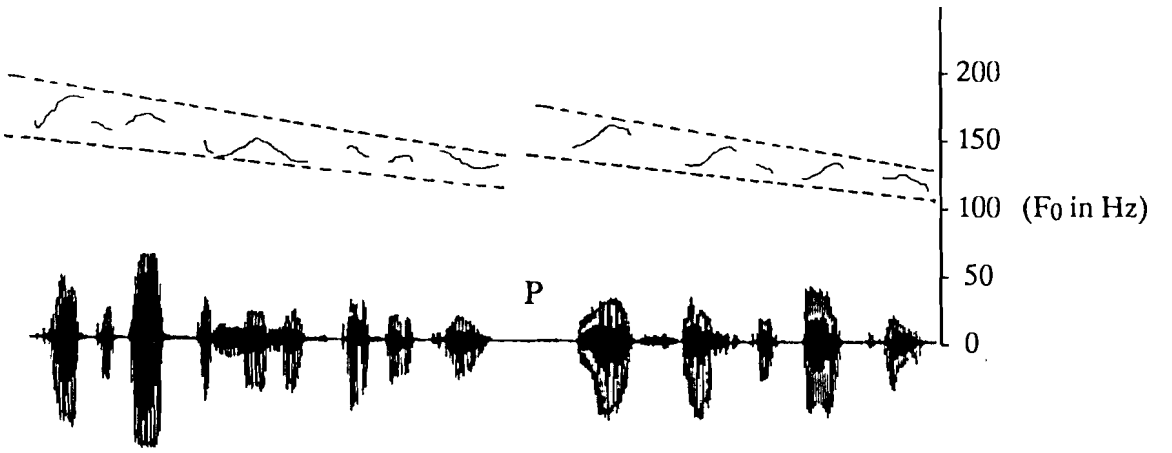
This chapter is organized as follows: Section 4.2 discusses the properties of resetting of F_0 contour across syntactic boundaries. In Section 4.3, we discuss the major factors which can affect the resetting of F_0 contour. The significance of pause between words, intonational phrases and sentences are discussed in Section 4.4.

4.2 Resetting of F_0 contour across syntactic boundaries

From our experiments on read sentences of complex and compound sentences, we have observed certain features related to the resetting across syntactic boundaries. For these studies we used a corpus of 100 sentences each with two syntactic clauses. Each syntactic clause forms an intonational phrase. The speech data of corresponding sentences are collected from three adult male native speakers of Hindi. The general properties of resetting of F_0 contour obtained from this analysis are summarized in following sections.

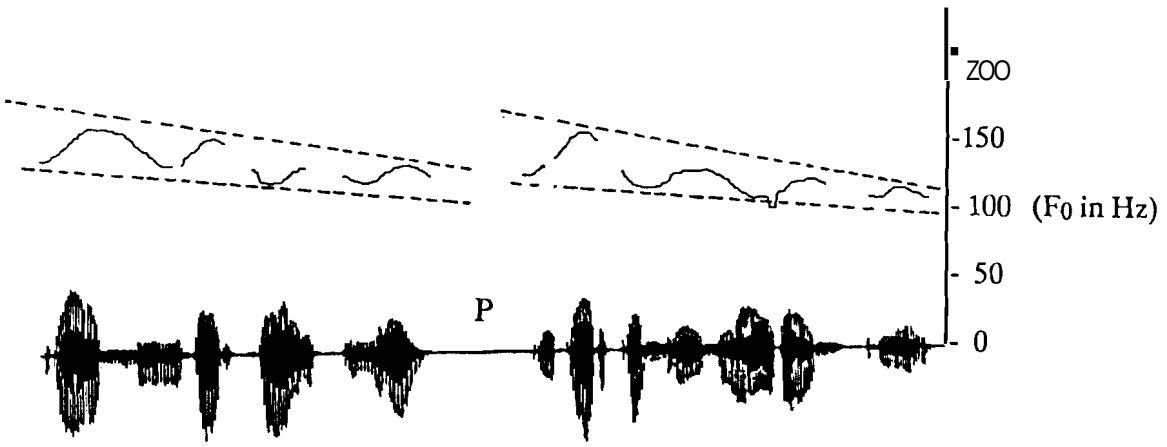
The initial peak F_0 (value of F_0 of the first peak of the first intonational phrase) is constant for a particular speaker. All other significant peaks and valleys in the subsequent clauses can be related to the initial peak Fo. The effect of resetting is directly proportional to the strength of syntactic boundary (Cooper & Paccia-Cooper, 1980). There is no significant resetting between short phrases. Within each syntactic clause, the F_0 contour exhibits declination tendency accompanied by local falls and rises.

Fig.4.1 shows the effect of resetting of F_0 contour across syntactic boundaries. **Fig.4.1a** is the speech waveform and corresponding Fo contour for the complex sentence */jo: kita:b* tumne: mujhko: *di:* thi: *vah ʃya:m* ke: *pa:s hai/* (The book which you gave me is with Syam). In figure, the F_0 which sets off (about 125 Hz) from onset of periodicity of the signal assumes maximum Fo level (about 180 Hz) within the same syllable for the first content word **(/jo:/).** The Fo contour drifts down from this point towards the initial syllable of the next content word **(/ki-/** in */kita:b/)* to about 130 Hz. Again, it rises towards a higher point (about 160 Hz) in the final syllable **(/-tab/** in */kita:b/)** of the word. The sentence */jo: kita:b* tumne: mujhko: di: thi: *vah ʃya:m* ke: *pa:s hai/* has two syntactic clauses and the F_0 contour drifts down as a function of time till the occurrence of major syntactic break (at the end of */jo: kita:b* tumne: mujhko: *di: thi:/),* which is also marked by a significant pause of duration of about 300 **msec.** That is, the utterance has two intonational phrases separated by a pause. The F_0 contour shows a similar behavior in the second intonational phrase also. **Fig.4.1b** is the



j o: k i t a: b tu mne: mujhko: di: th i: v a h ʂ y a: m k e: p a: s h ai

(a)



ra: m ne: idha r ʧ a: y p i ya: u dh a r ka mr e: mẽ: a:ra:m k i ya:

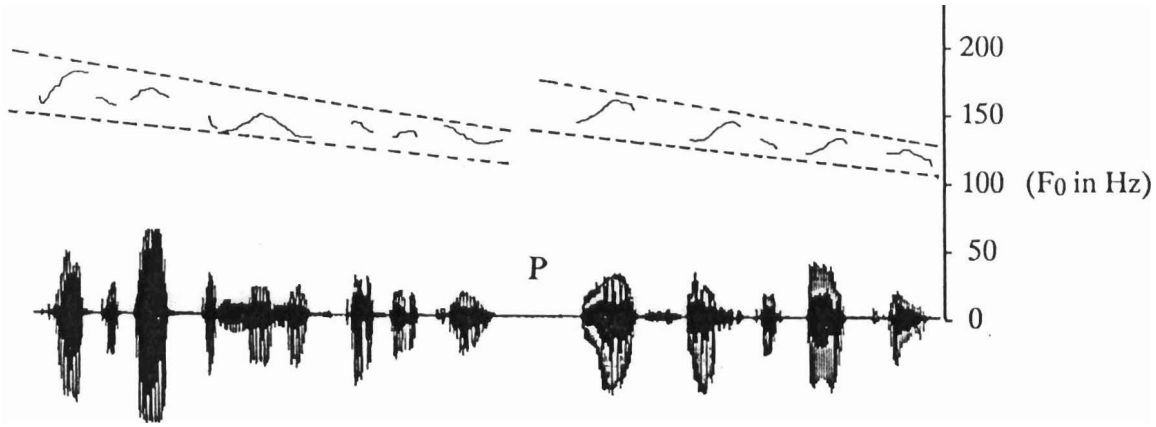
(b)

Fig.4.1. Speech waveform and F_0 contour for complex declarative sentences

(a) */jo: kita:b tumne: mujhko: di: thi: vah śya:m ke: pa:s hail* (The book which you gave me is with Syam)

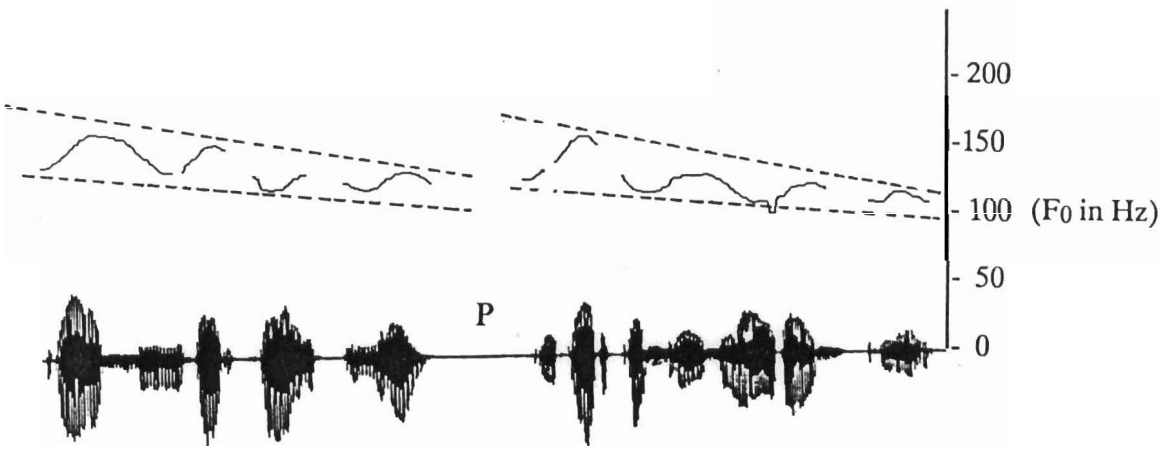
(b) */ra:m ne: idhar ča:y piya: udhar kamre: mẽ: a:ra:m kiya:/* (Ram took tea from here and took rest in that room)

Each sentence has two syntactic clauses separated by a pause (P). Resetting of F_0 contour occurs at the beginning of a new syntactic clause.



j o: k i t a: b tu mne: mujhko: di: thi: v a h ʃ y a: m k e: p a: s h a i

(a)



r a: m ne: idha r č a: y p i ya: u dh a r ka mr e: mẽ: a:ra:m k i ya:

(b)

Fig.4.1. Speech waveform and F_0 contour for complex declarative sentences

(a) */jo: kita:b tumne: mujhko: di: thi: vah śya:m ke: pa:s hai/* (The book which you gave me is with Syam)

(b) */ra:m ne: idhar ča:y piya: udhar kamre: mẽ: a:ra:m kiya:/* (Ram took tea from here and took rest in that room)

Each sentence has two syntactic clauses separated by a pause (P). Resetting of F_0 contour occurs at the beginning of a new syntactic clause.

natural utterance and Fo contour for the sentence */ra:m ne: idhar ca:y piya: udhar kamre: mẽ: a:ra:m kiya:/* (Ram took tea from here and took rest in that room). As like in Fig.4.1a, this sentence also has two intonational phrases and hence one major syntactic boundary (at the end of */ra:m ne: idhar ca:y piya/).* The resetting of F_0 contour coincides with this syntactic boundary.

4.3 Factors affecting the resetting of F_0 contour

The major factors which can affect the resetting of F_0 contour are physiological constraints, syntactic constraints and semantic constraints. Physiological constraints are the limitations imposed by speech production mechanism. Syntactic constraints include the change in resetting of F_0 contour with respect to the changes in the type of the sentence. Semantic constraints are the semantic aspects which control the properties of resetting of F_0 contour. In the following sections we discuss each of these constraints in detail.

43.1 Physiological constraints

Physiologically pitch frequency resetting can be explained in terms of *breathgroup* concept. A breathgroup is defined as the speech output that results from the synchronized activity of the chest, abdominal and laryngeal muscles during the course of a single expiration (Lieberman, 1967). The breath group sets the limit for any declarative sentence. However, when the sentence is long, pauses are given at the major syntactic boundaries. During a pause the subglottal pressure is built-up again and this is characterized by the resetting of F_0 contour.

There exist two physiological constraints on the resetting of F_0 contour. They are: 1) the built-up of subglottal air pressure during a pause and 2) the drop in subglottal air pressure during phonation. The corresponding acoustic parameters are pause between intonational phrases and duration of previous intonational phrase. We discuss each of these effects in the following sections.

4.3.1.1 Effect of pause between intonational phrases on resetting of F_0 contour

There is a correlation between pause at syntactic boundary and resetting

value. The stronger the boundary, the longer the pause. When the pause is long, subglottal pressure is built-up again and the F_0 contour resets to a high value. The value of resetting of F_0 contour increases with the pause up to a limit. Beyond the limit the pause does not show any correlation with the value of resetting. In such cases each syntactic clause can be considered as an independent sentence. From our experiments, we have observed that the pause between intonational phrases varies from 9% to 18% of the total duration of the utterance. However, the exact value of pause is determined by various syntactic constraints discussed in Section 4.3.2.

4.3.1.2 Effect of duration of previous intonational phrase on resetting of F_0 contour

Like pause, duration of the previous intonational phrase also has an effect on the resetting of F_0 contour. If the pause remains unchanged, the value of resetting of F_0 contour will be less after a long intonational phrase. This is due to the inability of the human speech production mechanism to restore sufficient subglottal air pressure after a long intonational phrase. Thus the value of 'resetting of F_0 contour is inversely proportional to the length of the previous syntactic clause.

43.2 Syntactic constraints

The syntactic construction of sentences affects the resetting of F_0 contour significantly. In Hindi, resetting of F_0 contour occurs in complex declarative and compound sentences. The value of resetting of F_0 contour and the pause between intonational phrases vary with different types of sentences. In the following sections we discuss the effect of resetting of F_0 contour in complex declarative sentences and compound sentences.

4.3.2.1 Effect of resetting of F_0 contour in complex declarative sentences

Sentences in which a main clause occurs with one or more subordinate clauses are called complex declarative sentences. These subordinate clauses are determined by subordinate conjunctions in Hindi (Kachru, 1980). The

subordinate conjunctions in Hindi include relative-correlative clause, complement clause, purpose clause, reason clause, etc.. Different subordinate conjunctions and corresponding keywords in Hindi are shown in Table 4.1. The effect of resetting is separately studied for relative-correlative clause and

Conjunctions	Keywords
1. relative-correlativeclause	
relative clause	jo-vah
when clause	jab-tab
as long as	*jab* tak-tab tak
where	*jahã:-vahã:*
as far as	*jahã: tak-vahã:* tak
which direction	jidhar-udhar
which manner	jake:-vaise:
which quality	*jaisa:-vaisa:*
which quantity	*jitna:-utna:*
2. compliment clause: that	ki
3. purpose clause	*isliye:* (ki), taki
4. reason clause	*kyõ:ki:, cu:ki-isiliye:*
5. concessive clause	yadycpi-to: bhi:
	ha:la:ki-phir bhi:
6. conditional clause	*yadi,* agar-to:
7. contradictory clase	cahe:
8. result clause	*isi:liye:, atah, the:va:*
9. otherwise	*anyatha:*
10. apprehension	kahi: ....*na*

Table **4.1.** Subordinate conjunctions in Hindi (Kachru, **1980)**

nonrelative clauses. Following are the conclusions made from this study.

4.3.2.1.1 Relative-correlativeclauses in Hindi

Relative-correlative clause can be divided into different subclauses like relative, when, as long as, where, as far as, which direction, which quality, which quantity, etc.. Each of these clauses can be separated from a text by the presence of the keywords */jo:-vah/,* Jab-tab/, */jidhar-udhar/,* etc. as shown in Table **4.1.** Structurally relative-correlative clause is similar to antecedent-consequence form. The first part of the sentence which contains the first element in the keyword pair can be treated as the antecedent. The second part of the sentence starts from the second element of the keyword pair and ends with the end of the sentence. Hence by using this the keyword pair we can divide the sentence into two syntactic clauses and the resetting of F_0 contour occurs at the beginning of the second syntactic clause. During speaking, the two intonational phrases (corresponding to two syntactic clauses) are separated by a pause. To show the properties of intonation patterns in relative clause of complex declarative sentences, we had selected 10 such sentences (list of sentences are given in Appendix B4) and speech data were collected from three adult male native speakers. Results from this study is discussed below.

In order to generalize the results, we normalized durational information with respect to the total duration of the utterance and the values of Fo were normalized to the initial peak F_0 (value of F_0 of the first peak of the first intonational phrase). The results discussed below are based on the average values of three speakers. The individual differences in normalized values are negligible.

Variation in the duration of the first intonational phrase was from **35.26%** to **59.26%** of the total duration of the utterance. We have observed that the pause between the clauses is around **12.27%** (ranging from **10.92%** to **13.27%** with a standard deviation of **0.85)** of the total duration of the utterance.

The average initial peak F_0 is around **180** Hz for an adult male speaker and is speaker dependent (It was **170** Hz, **182** Hz and **188** Hz for three speakers we

tried). All other significant peaks such as the intermediate and the final peaks of the intonational phrases, resetting value of F_0 (the first peak of the second intonational phrase) and tapering frequency at the end of the utterance can be related to the initial peak frequency. The final peak of the first intonational phrase is about 71.58% (ranging from 68.34% to 75.42% with a standard deviation of 2.28) of the initial frequency. Resetting frequency is around 90.83% (ranging from 87.66% to 94.67% with a standard deviation of 2.05) and the final peak of the second intonational phrase is around 69.48% (ranging from 65.25% to 72.25% with a standard deviation of 2.30) of the initial peak frequency. End tapering frequency is always constant, that is, around 56% of the initial peak frequency. The detailed experimental results for the resetting of peaks are given in Appendix C13.

Like peaks, valleys in complex declarative sentences also show a systematic behavior. We have analyzed all significant valley points with respect to the initial peak frequency. After resetting, the values of the valleys also change similar to the values of the peaks. From the analysis we have observed that generally the initial valley (the valley of first prosodic word in the first intonational phrase) stands out separately from the base line. Its value is around 69.72% (ranging from 62.15% to 73.65% with a standard deviation of 3.06) of the initial peak frequency. The base line of the first intonational phrase is the line joining the second valley (valley of the second prosodic word in the first intonational phrase) and the final valley of the first intonational phrase. In the first intonational phrase, the second valley is around 75.37% (ranging from 70.29% to 79.27% with a standard deviation of 2.91) and the final valley is 63.07% (ranging from 60.15% to 69.25% with a standard deviation of 2.60) of the initial peak frequency. The base line of the second intonational phrase is the line joining the resetting valley and the final valley of the second intonational phrase. The resetting valley is around 68.81% (ranging from 65.23% to 74.34% with a standard deviation of 2.35) and the final valley of the second intonational phrase is around 62% (ranging from 59.27% to 65% with a standard deviation of 1.95) of the initial peak frequency. Appendix C14 gives the results obtained from the analysis of

valleys in complex declarative sentences.

4.3.2.1.2 *Complex* declarative sentences of nonrelative clauses

Complex declarative sentences include clauses other than relative-correlative clause such as complement, purpose, reason, concessive, conditional, contradictory, result, otherwise and apprehension clauses as shown in Table 4.1. Keywords for identifying these clauses from a text are */ki/, /isliye/, /kyõ:ki/, /yadya:pi-to: bhi:/*, etc.. The syntactic clauses can be distinguished by the presence of keyword or the second element if keywords are in pair. F_0 contour resets at the keyword and pause is inserted before the keyword. Following are the results of the experiments conducted to analyze the resetting of F_0 contour of complex declarative sentences of nonrelative clause. 10 sentences were selected (sentences are given in Appendix B5) and corresponding speech data were collected from three adult male native speakers of Hindi. The results discussed below is the normalized average value computed as mentioned in the previous case. Among speakers, difference in normalized values were very small.

In general, we found that for nonrelative clause of complex declarative sentences in Hindi the value of resetting of F_0 contour and the amount of pause between intonational phrases are slightly lesser than the relative clause. All sentences have two intonational phrases in which the duration of the first intonational phrase was from 33.14% to 57.98% of the total duration of the utterance. Pause between intonational phrases is around 10.88% (ranging from 9.56% to 12.67% with a standard deviation of 1.05) of the total duration of the utterance. Following are some of the conclusions about peak and valley resetting of F_0 contours of nonrelative clause of complex declarative sentences.

As in the previous case all significant peaks and valleys can be related to the initial peak F_0. The final peak of the first syntactic clause is around 69.96% (ranging from 67.20% to 73.35% with a standard deviation of 2.13) of the initial peak Fo. Resetting F_0 is around 88.02% (ranging from 82.81% to 93.95% with a standard deviation of 3.17) and the final peak of the second syntactic clause is around 68.20% (ranging from 64.67% to 72.12% with a standard deviation of

2.26). End tapering frequency is **55.48%** of the initial peak Fo.

The observations in the resetting of valleys also stand closer to the previous observation. The initial valley stands out separately from the base line and its value is around **70.44%** (ranging from **67.76%** to **72.32%** with a standard deviation of **1.43)** of the initial peak F_0. In the first syntactic clause the second valley is around **73.93%** (ranging from **69.34%** to **76.54%** with a standard deviation of **2.57)** and the final valley is **61.34%** (ranging from **59.23%** to **65.43%** with a standard deviation of **1.57)** of the initial peak F_0. The resetting valley is around **67.94%** (ranging from **63.260%** to **72.12%** with a standard deviation of **2.92)** and the final valley of the second syntactic clause is around **59.19%** (ranging from **56.34%** to **63.80%** with a standard deviation of **2.28)** of the initial peak Fo.

Appendices **C15** and **C16** show the results of the peak and valley resetting for nonrelative clause of complex declarative sentences in Hindi. From these observations, we can conclude that resetting of F_0 contour in relative and nonrelative clauses of complex declarative sentences show **similar** behavior. But in the relative clause, syntactic clauses are more stronger and hence the value of resetting of F_0 contour and the pause between the clauses are also more. In relative clause, pause between intonation phrases is about **12.27%** of the total duration of the utterance and the value of resetting for peaks of F_0 is **90.83%** of the initial peak F_0. In nonrelative clause these measurements are **10.88%** and **88.02%,** respectively.

4.3.2.2 Effect of resetting of F_0 contour in compound sentences

Sentences in which two or more independent clauses are joined together with coordinate conjunctions are called compound sentences (Kachru, **1980).** The coordinate conjunctions in Hindi can be conjunctions, disjunctions, negative disjunctions or adversative conjunctions. Keywords belonging to this clauses are */aur/, /va:/, /evam/, /tatha:/,***etc..** The detailed list of different classes of coordinate conjunctions and corresponding keywords are given in Table **4.2.** The conjunctions like */aur/* (and) may occur in between two noun phrases also. So **identifying** syntactic clauses from a given sentence need some more higher level

Cojunctions	Keywords
1. conjunction: and	*aur, v a , e:vam, tatha:*
2. disjunction: or	*ya:, v a , athva, kimva:*
3. negative disjunction: neither...nor	*na....na*
4. adversative conjunction: but	*par, parantu:, kintu:, le:kin, magar*

Table 4.2. Coordinate conjunctions in Hindi (Kachru, 1980)

knowledge. Cues like the total number of words in a syntactic clause, word just before the conjunction, etc. can be used to separate the syntactic clauses in the sentence. For example, the sentence */sudha:kar bahu:t cãcal hai aur sudhi:r bahu:t ś a:nt hail* (Sudhakar is very restless and Sudhir is very calm) can be divided into two syntactic clauses (*/sudha:kar bahu:t cãcal hail* (Sudhakar is very restless) and */sudhi:r bahu:t ś a:nt hail* (Sudhir is very calm)) by make use of the presence of the function word */hai/* (is) just before the coordinate conjunction */aur/* (and). To show the behavior of intonation pattern for compound sentences in Hindi we have considered 10 sentences (sentences are given in appendix B6) and the corresponding speech data were collected from three adult male native speakers of Hindi. For the analysis the values were normalized and averaged with respect to the total duration and the initial peak F_0. The results from this analysis are shown below.

All sentences had two intonational phrases in which duration of the first intonational phrase varies from 36.18% to 57.12% of the total duration. Pause between two syntactic clauses is around 14.24% (ranging from 12.58% to 17.22% with a standard deviation of 1.31) of the total duration of the utterance. The final peak F_0 of the first syntactic clause is about 79% (ranging from 70.18% to 83.28% with a standard deviation of 3.71) of the initial peak Fo. Resetting Fo is around 93.08% (ranging from 90.88% to 96.82% with a standard deviation of 2.25) and

the final peak F_0 of the second syntactic clause is around 70.01% (ranging from 65.32% to 74.40% with a standard deviation of 2.21) of the initial peak Fo. End tapering frequency is constant and is around 56% of the initial peak Fo.

Like resetting in the peaks, valleys also get modified across syntactic boundaries. As in the previous cases, here also the initial valley stands out separately from the base line and its value is around 71.18% (ranging from 67.69% to 74.26% with a standard deviation of 1.91) of the initial peak Fo. In the first syntactic clause, the second valley is around 81.04% (ranging from 79.26% to 83.57% with a standard deviation of 1.26) and the final valley of the first syntactic clause is around 67.11% (ranging from 65.01% to 70.78% with a standard deviation of 2.43) of the initial peak F_0. The resetting valley is around 74.28% (ranging from 72.13% to 77.23% with a standard deviation of 1.80) and the final valley of the second syntactic clause is around 66.04% (ranging from 60.14% to 70.56% with a standard deviation of 3.32) of the initial peak F_0. Appendices C17 and C18 summarize the results obtained from the analysis of valleys and peaks in compound declarative sentences, respectively.

Compound sentences are formed from two independent clauses using coordinate conjunctions. The strength of each clause is more or less the same. This is reflected as the increase in the values of resetting of F_0 contour and the pause between intonational phrases. But these values are not very large from that of complex declarative sentences. For example, for compound sentences the value of resetting of F_0 contour at peaks is 93.08% and pause between intonational phrases is 14.24%. But for the relative clause of complex declarative sentences corresponding values are 90.83% and 12.27%.

4.3.2.3 Effects of resetting in sentences with more than two syntactic clauses

F_0 contour will reset at the beginning of each intonational phrase. Each resetting is preceded by a proper amount of pause to restore the subglottal air pressure. So far we considered sentences with two syntactic clauses. These sentences contains two intonational phrases separated by a pause. If a sentence contains more than two intonational phrases F_0 resets at the beginning of each

intonational phrase and pause is inserted between the phrases. From experiments, we found that if the sentence has more than two syntactic clauses the pauses between the clauses are less. Correspondingly, the value of resetting of F_0 contour also decreases. But the magnitude of decrease is very less.

In all the cases of syntactic constraints we have considered the initial peak F_0 as the reference point. Hence the accuracy of modeling of the resetting of Fo contour is depends on the choice of the initial peak value of F_0 contour.

4 3 3 Semantic constraints

Like physiologic and syntactic constraints, semantic aspects of sentences play a role in the resetting of F_0 contour. F_0 contour changes significantly if the speaker gives more emphasis for selected words during discourse. Hence emphasis is one of the major factor which affects the F_0 contour. Other prosodic properties like duration and gain are also affected by emphasis. Experiments on emphasis are not included in the scope of this research work since we are mostly concentrated on the intonational behavior of neutral declarative sentences. In the following sections we discuss the changes of F_0 contour due to two of the commonly occurring semantic issues in any Hindi text. They are the effect of negation and the effect of numerals.

4.3.3.1 Effect of negation

Negation in Hindi sentences is signaled by the presence of the function words */nahĩ:/* or */na/*. The presence of these words compliment the meaning of the sentence itself. For example, consider the sentences */mãĩ ka:ɳpur ja: raha: hũ:/* (I am going to Kanpur) and */mãĩ ka:ɳpur nahĩ: ja: raha: hũ:/* (I am not going to Kanpur). The meaning of the second sentence compliments the meaning of the first one by the addition of the word */nahĩ:/*. In order to analyze the effect of F_0 change during negation we constructed 10 sample sentences each considered with and without negation (sentences are given in Appendix B8) and the speech data were collected from three adult male native speakers of Hindi. From this analysis we found that during negation the value of F_0 peak rises to 15 to 20 Hz than the top line. Similar changes occur in the valley of the negation word.

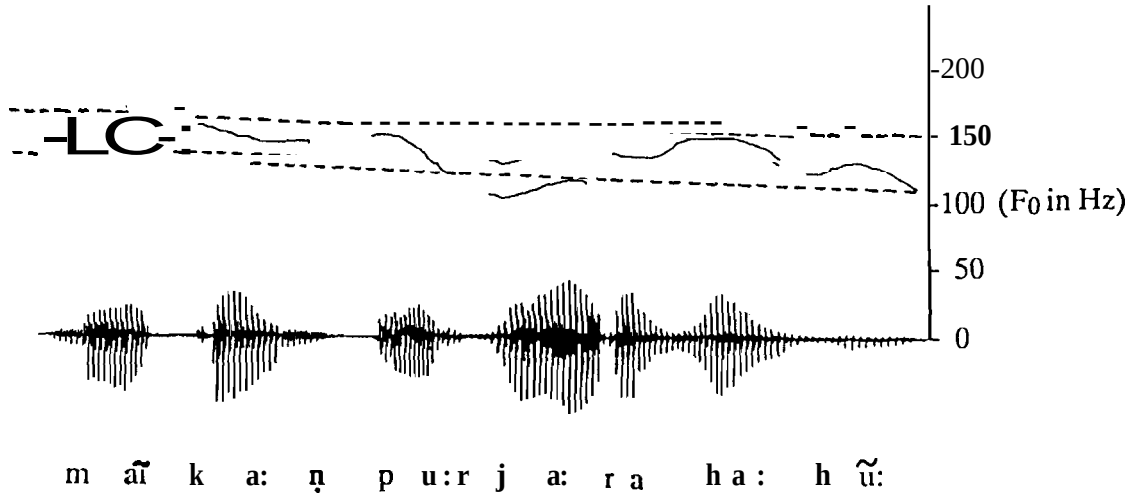
These are due to the emphasis given by the speaker during negation. Significantly, this change in F_0 value did not accompanied by the presence of significant pause.

Fig.4.2 shows the change in F_0 contour due to the presence of negation. Fig4.2a is the speech waveform and the corresponding F_0 contour for the utterance */maĩ ka:ɳpur ja: raha: hũ:/* (I am going to Kanpur) . As shown in the figure the valleys and peaks of F_0 contour lie within the limits of the top line and the base line. Fig.4.2b is the speech waveform and corresponding F_0 contour for the negated utterance */maĩ ka:ɳpur nahĩ: ja: raha: hũ:/* (I am not going to Kanpur). In Fig.4.2b it can be seen that the valley and peak of the negation word */nahĩ:/* (not) will not match with the top line and the base line. The peak of the word */nahĩ:/* is nearly 20 Hz higher than the normal peak at that point and valley is nearly 15 Hz higher than the normal valley at that point. The values of the valleys and peaks are changed because of the emphasis given by the speaker to convey the semantic change by negation.

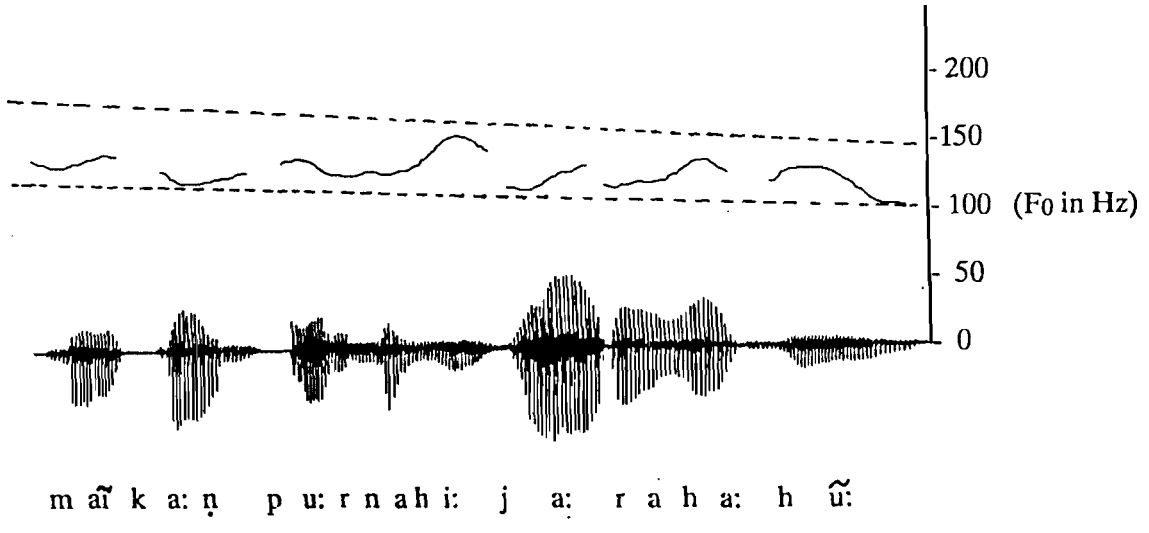
4.3.3.2 Effect of numerals

Numerals occur very frequently in any text. While reading a text the speaker has to expand the numerals into the corresponding spoken form and then utter. Altogether, the presence of numerals gives a different semantic interpretation and hence the speaker gives more emphasis on that. In order to find out the behavior of F_0 contour by the presence of numerals we have analyzed 10 sentences uttered by three adult male native speakers of Hindi (sentences are given in Appendix B9). From the experiments we noticed that the valleys and peaks of F_0 contour reset to a new value at the first word .(the numeral in the left most position) in a sequence of numerals. This effect is very small if the spoken form of the numeral contains less number of words. The resetting of F_0 contour due to the presence of numerals did not accompanied by significant amount of pause.

Fig.4.3 shows the effect of numerals in Hindi. From the figures it is obvious that the valley and the peak values of F_0 contour reset at the beginning of the



(a)



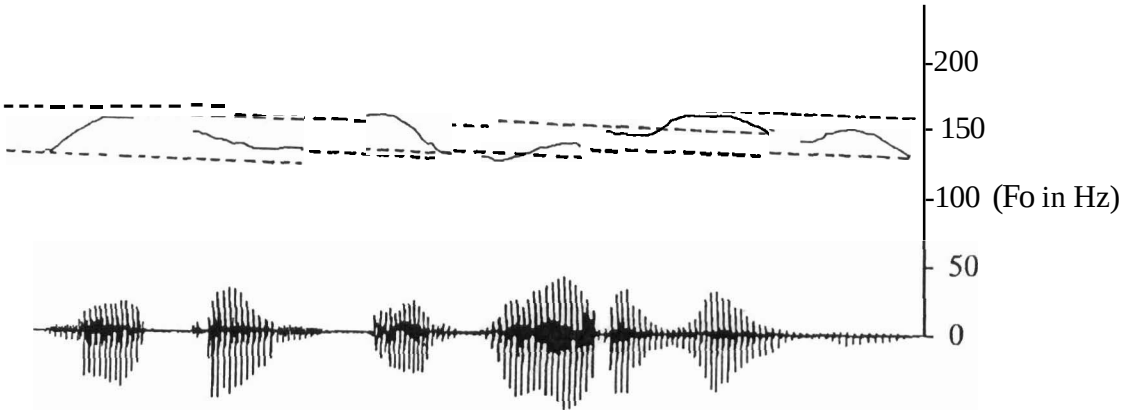
(b)

Fig.4.2. Changes of F_0 contour due to the presence of negation

(a) F_0 contour and the utterance for the sentence */maĩ ka:ṇpu:r ja: raha: hũ:/* *(I* **am going to Kanpur)**

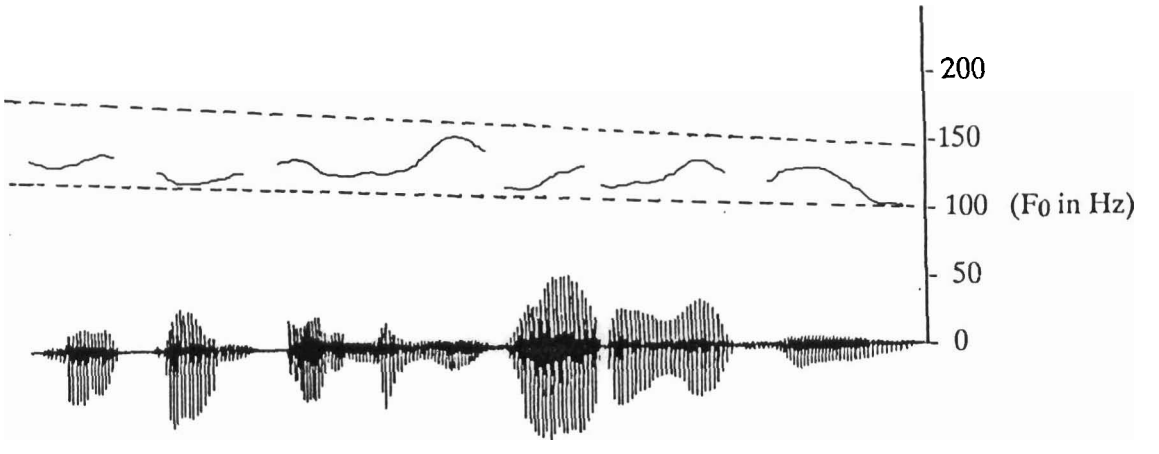
(b) F_0 contour and the utterance for the sentence */maĩ ka:ṇpu:r nahĩ: ja: raha: hũ:/* *(I* **am not going to Kanpur)**

The valleys and the peaks F_0 contour in the negation word */nahĩ:/* **is emphasized in** (b).



m aĩ k a: ṇ p u: r j a: r a h a : h ũ:

(a)



m aĩ k a: ṇ p u: r n a h i: j a: r a h a: h ũ:

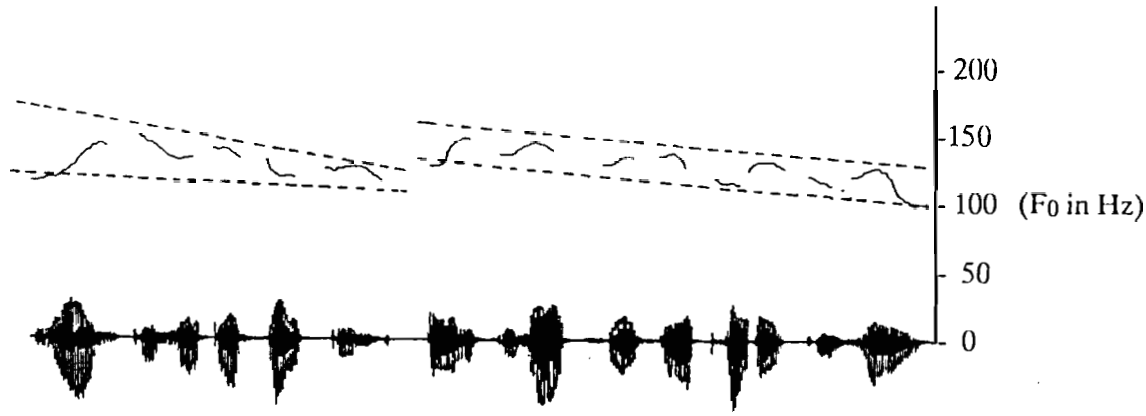
(b)

Fig.4.2. Changes of F_0 contour due to the presence of negation

(a) F_0 contour and the utterance for the sentence */maĩ ka:ṇpu:r ja: raha: hũ:/* (I am going to Kanpur)

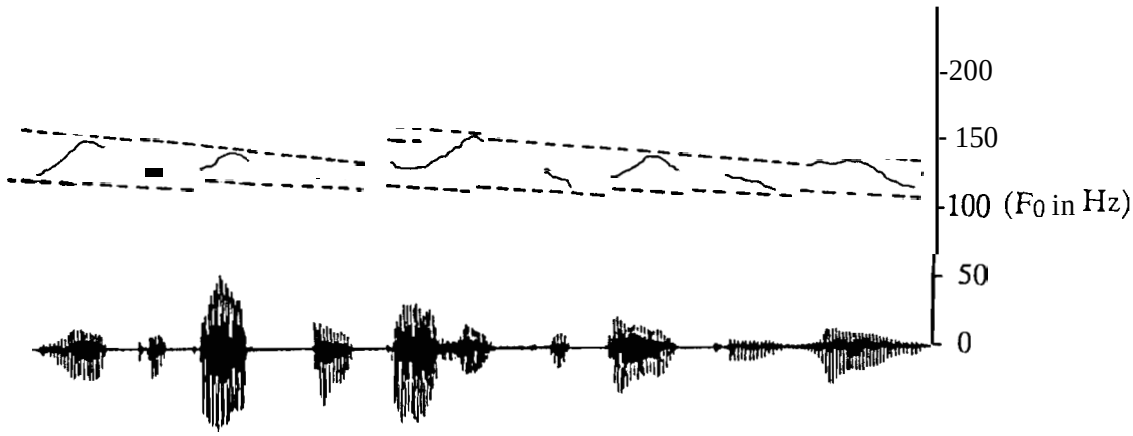
(b) F_0 contour and the utterance for the sentence */maĩ ka:ṇpu:r nahĩ: ja: raha: hũ:/* (I am not going to Kanpur)

The valleys and the peaks F_0 contour in the negation word */nahĩ:/* is emphasized in (b).



r a: m ki: mo:ţar s a:i: k i l t e: i: s haj a:r ch e: s auaṭha:rah k i: h ai

(a)



y e: k i t a: b painta: l i s rupaye: k i: h ai

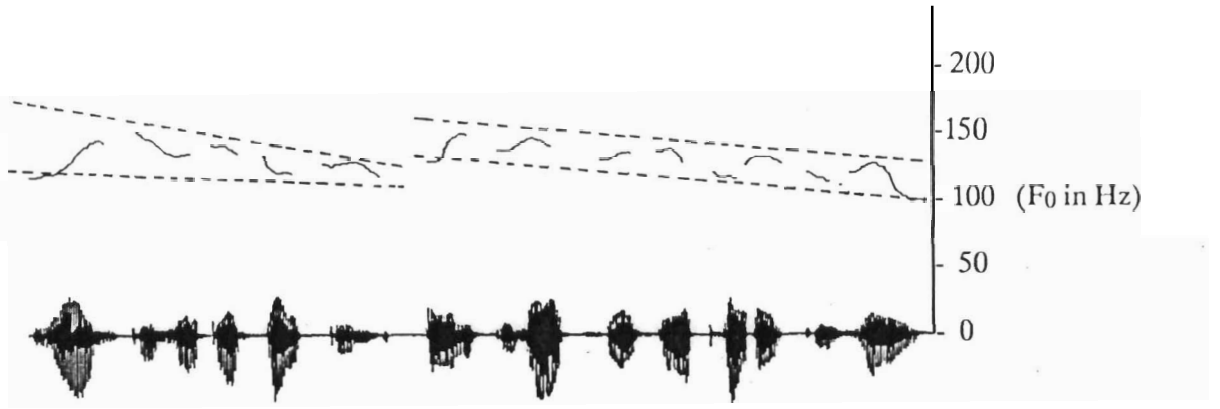
(b)

Fig.4.3. Changes of F_0 contour due to the presence of numerals

(a) */ra:m ki: mo:ţar sa:i:kil te:i:s haja:r che: sau aṭha:rah ki: hail* (Ram's motor cycle costs twenty three thousand six hundred and eighteen.

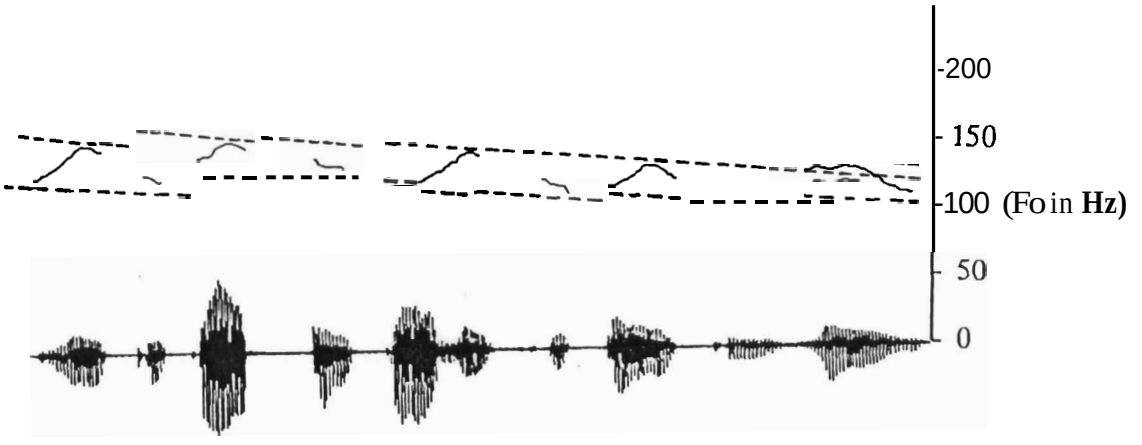
(b) */ye: kita:b paintha:lis rupaye: ki: hail* (This book costs forty five rupees)

F_0 contour resets at the beginning of the numeral and the magnitude of resetting is proportional to the number of words in the numeral.



r a: m ki: mo:ţar s a:i: k i l t e: i: s haj a:r ch e: s auaţha:rah k i: h ai

(a)



y e: k i t a: b painta: l i s rupaye: k i: h ai

(b)

Fig.4.3. Changes of F0 contour due to the presence of numerals

(a) */ra:m ki: mo:ţar sa:i:kil te:i:s haja:r che: sau aţha:rah ki: hail* (Ram's motor cycle costs twenty three thousand six hundred and eighteen.

(b) *lye: kita:b paintha:lis rupaye: ki: hail* (This book costs forty five rupees)

F0 contour resets at the beginning of the numeral and the magnitude of resetting is proportional to the number of words in the numeral.

numeral with out the presence of significant amount of pause. In Fig.4.3a resetting of F_0 contour occurs at the beginning of third word which corresponds to the left most numeral in the text. Consider the case of Fig.4.3b. Here the spoken form of numeral consists of lesser number of words and correspondingly the amount of resetting is small. In both the cases the presence of numerals affect the peaks and valleys of F_0 contour of the subsequent words.

4.4 Pause: The sound of silence

Pauses have been assigned to two main functions: 1) they separate large grammatical units, such as syntactic clauses and 2) they serve to clarify subgrouping of smaller units (Kutik, Cooper, Boyce, 1983). The major function of a pause is to demarcate a complex sentence into different intonational phrases. We already discussed the factors which controls the duration of a pause between intonational phrases. Pauses occur between words, intonational phrases and sentences. In the following section we discuss each of these occurrences in detail.

4.4.1 Pause between words

Speakers are giving different durations of pauses between words while uttering a continuous text. The duration of pause between words is controlled by different features like lexical content of the words, position of the word in an intonational phrase and the phonetic behavior of pre-pause and post-pause syllables.

Before the words of high lexical content, the amount of pause is slightly more. In normal speech, the word preceded by a pause is often difficult to guess in advance. This sort of pause typically occurs before a minor constituent boundary, generally before a noun phrase, verb phrase or adverbial phrase. Also if the number of syllables in a word is more (greater than three) then the amount of pause before the word is more. The effect of lexical content on pause can be considered as the difficulty in finding the next word in an utterance.

Position of a word in an intonational phrase also carries some weight to compute the pause. In normal discourse the pause after the first word in an intonational phrase is more. Also, towards the end of intonational phrase the

pauses between words are less. Pauses due to positional effect usually occur after the first word in an intonational phrase. It seems to serve a planning function. It is essentially a holding operation while the speaker plans the remainder of the sentence. Pauses due to the lexical content and the position do not occur before the most prominent word (usually the first content word) in an intonational phrase.

Phonetic behavior of pre-pause and post-pause syllables have some effect on deciding the duration of the pause between the words. From our experiments, we noticed that if pre-pause syllable ends and post-pause syllable starts with vowels then the duration of corresponding pause is more. For example in the intonational phrase */a:tma: amar* hail (Soul is immortal), the final syllable of the first word is */-ma:/* and the first syllable of the second word is */a-/* and hence the duration of the pause between the first word and the second word is more than the pause between the second word and the third word. This durational difference in pause changes with different vowel combinations. We found that the amount of pause decreases with the difference in the quality of vowels. For instance, the pause is the minimum if pre-pause and post-pause syllables are high and low vowels and the maximum if both vowels belong to the same quality.

Pauses between words can be taken as a minor feature internal to an intonational phrase. They are not considered as the markers of major phrases in an utterance because they do not result in utterance chunks each of which has an Fo contour and typically contained within an intonational phrase. The amount of pause between words is much less than the amount of pause between intonational phrases.

4.43 Pause between intonational phrases

As discussed earlier, intonational phrases are delimited by a significant amount of pause and resetting of F_0 contour. There is a correlation between the type of constituent boundary and the length of the pause. That is, the more major the boundary, the longer the pause. The amount of pause between intonational phrases can be related to the total duration of the utterance as discussed in

Section **4.3.2.**

We discussed the effect of pause between intonational phrases with respect to the syntactic variations of the sentence. The pause between the intonational phrases in compound sentences are the maximum **(14.24%** of the total duration of the utterance). Among complex declarative sentences, the relative clause exhibits more amount of pause between intonational phrases **(12.27%** of the total duration of the utterance) than nonrelative clause of sentences **(10.88%** of the total duration of the utterance). This type of pauses in an utterance can generally be taken as a cue for phrase boundary.

4.43 Pause between sentences

While reading a paragraph text we have to give proper amount of pause between sentences. The amount of such a pause is maximum than the pause between words and the pause between intonational phrases. The duration of pause between sentences is controlled by several factors like duration of the sentence spoken, topic change between sentences, etc.. Longer pause is given between sentences if the sentences are long. It is for the full restoration of the drop in subglottal air pressure. Also pause between sentences tend to be longer if two successive sentences discuss different topics.

4.5 Summary

In this chapter we have discussed the properties of intonation patterns for complex declarative and compound sentences in Hindi. F_0 contours for such sentences reset across major syntactic boundaries. Resetting of F_0 contour is affected by several constraints imposed by physiological, syntactic and semantic factors. These issues were discussed in detail. The resetting of F_0 contour is also accompanied by a significant amount of pause. Pause between words, intonational phrases and sentences were also discussed.

Acoustic-phonetic properties of the constituent speech units alter the properties of intonation patterns slightly. In the following chapter we discuss the effects of segmental factors on F_0 contour of an utterance.

Chapter 5
EFFECTS OF SEGMENTAL FACTORS ON F_0 CONTOUR

5.1 Introduction

Acoustic-phonetic properties of speech sounds alter the properties of intonation patterns to some extent. But the influences of these properties are less when compared with the global and local properties of F_0 contour discussed in previous chapters. The segmental factors are determined by the constraints of human speech production mechanism and hence these are considered to be language independent. This chapter discussed the effects of segmental factors on F_0 contour for continuous speech in Hindi.

This chapter is organized as follows: Section 5.2 discusses the acoustic-phonetic properties of speech sounds. It includes discussion on the classification of vowel and consonant speech sounds. Experimental set up used for the analysis of segmental properties is discussed in Section 5.3. Section 5.4 discusses the properties of inherent F_0 and the effects of surrounding speech units on the inherent F_0 of the present vowel.

5.2 Acoustic-phonetics of speech sounds

Generally, speech sounds can be classified into consonants and vowels. Both in the vocal tract shaping and in the sound source consonants differ from vowels. The sound source for vowels is always periodic but a consonant may be produced with a periodic sound source, an aperiodic sound source or a combination of both periodic and aperiodic sound sources. In the following sections, we discuss the classification of vowel and consonant speech sounds based on their acoustic-phonetic behavior.

52.1 Classification of vowels

Hindi has a set of eight vowels. They are /a/, /a:/,/i/, /i:/, /u/, /u:/, /e:/ and /o:/. Among these /a/, /i/ and /u/ are the shorter versions of the long vowels /a:/, /i:/ and /u:/, respectively. So for the classification of the vowel sounds we can consider five long vowels such as /a:/, /i:/, /u:/, /e:/ and /o:/. These vowel sounds can be classified based on three terms. They are: (1) height of the body of the tongue; (2) front-back position of the tongue and (3) the degree of lip rounding (Catford, 1988; Ladefoged, 1975). This classification is done based on the observation of the tongue and mouth movements of the speaker during phonation. For the vowels /i:/ and /e:/, the highest point of tongue is the front of the mouth and hence called front vowel. The tongue is closed to upper or back surface of the vocal tract for the vowels /a:/, /o:/ and /u:/. They are called back vowels. Table 5.1 shows the classification of vowels in Hindi based on tongue height and position.

The degree of lip rounding can be predicted from the degree of backness and the degree of height of the tongue in most of the languages (Ladefoged, 1975; Lieberman & Blumstein, 1988). Front vowels are usually unrounded and back vowels are usually rounded. The degree of roundness increases with the degree of height. Hence by considering all these, vowels of a language can be kept in maximally distinct positions. For the classification of Hindi vowels, we made use of these assumptions.

52.2 Classification of consonants

Consonants can be classified into different categories based on their articulatory features. The acoustic features which correspond to each articulatory feature form the basis of perceptual distinction from one consonant to another. Hence each feature category is described by the joint consideration of three levels of linguistic analysis. They are: (1) perceptual; (2) acoustic and (3) linguistic levels. This analysis form a consistent frame work for specifying the communicative sound structure of any language (Chomsky & Halle, 1968).

The articulatory features of consonants are of three types. They are: (1) feature of manner of articulation, (2) the voicing feature and (3) feature of place

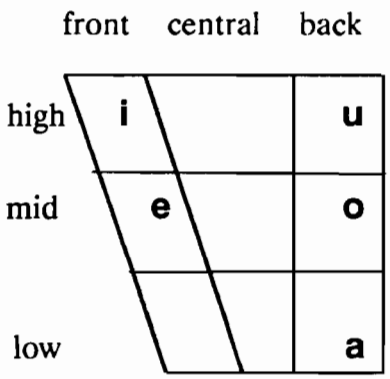


Table 5.1. Classification of vowels based on quality of the vowel and tongue position

		place of articulation						
manner of articulation		Bilabial	Dental	Alveolar	Retroflex	Palatel	Velar	glottal
	Stop	p b ph bh	t d th dh		ṭ ḍ ṭh ḍh		k g kh gh	
	Affricate					c j ch jh		
	Nasal	m	n		ṇ	ń	ṅ	
	Fricative			s	ṣ	ś		h
	Trill			r				
	Lateral			l				
	Glide		v			y		

Table 5.2. Classification of consonants based on place and manner of articulation

of articulation. The manner and voicing features are states of articulation irrespective of place of articulation. The voicing feature is either voicing or unvoicing of consonants which corresponds to the periodic or aperiodic sound sources. Manner features are stops, glides, nasals and fricatives. Glide articulation is classified again into semivowel, lateral and trill. Fricative sounds in Hindi are again classified into affricates and fricatives. Manner features differ in the constriction of oral tract. For example, stop consonant obstructs the breath stream completely during a portion of the articulatory gesture. Fricative consonants are formed with a narrow constriction and the semi vowel consonants constrict the oral tract more like high vowels. Articulatory place for consonants can be divided into front, middle and back. Labial and labio-dental belongs to front position, dental, alveolar and palatel are middle position consonants while velar and glottal are back position consonants. Table 5.2 shows the classification of Hindi consonants. Rows corresponds to place of articulation and columns corresponds to the voicing and manner of articulation.

5.2.3 Segmental constraints on inherent properties

The properties of a speech unit are affected by the properties of the surrounding speech units. This effect is significantly reflected in vowel speech sounds than consonants. For consonants, the production is little affected by the surroundings and the process of production is affected by following physiological factors. They are: (1) the constriction effect of consonant articulation of the oral tract; (2) the subglottal air pressure; (3) air pressure on the mouth and (4) the state of vocal folds (Fant, 1960; Flanagan, 1972; Heinz & Stevens, 1961; Picket, 1980; Kohler, 1990). These physiological factors reflect in the source parameters (F_0, gain and duration) of the following vowel unit. Similarly, the change of the following consonant and the preceding and the following syllables affect the values of F_0 of the vowel sound. We discuss each of these changes in section 5.4.

5.3 Experimental conditions

F_0 values of vowels were studied by embedding the test words in a carrier sentence */me:ra: na:m* _____ *hai/* (My name is _____). For this study we

have selected several possible combinations of disyllabic words. The test words are mostly nonsense words where the vowel characteristics were studied both the initial and the final syllable separately. In order to study the effect of consonants we constructed the carrier sentences with the test words of the form CV_1XV_2C where X is the position of the consonant under study. The change in the inherent F_0 of the vowel V_1 due to the change in the consonant X without altering any other character in the test word is classified as effect of the following consonant. Similarly the change in the inherent F_0 of the vowel V_2 due to the change in the consonant X is classified as effect of the preceding consonant. In order to analyze effect of the preceding syllable, we constructed test words of the form CXCVC. The change in the inherent F_0 of the vowel V due to the change in the nucleus of the previous syllable X without altering any other character is considered as effect of the preceding syllable. Similarly change in the inherent F_0 of the vowel V in a test word of type CVCXC due to the change in the nucleus of the following syllable X is considered as effect of the following syllable.

For analyzing the effect of segmental factors on F_0 contour we have collected speech data for 1700 nonsense disyllabic words embedded in the above carrier sentence from two adult male native speakers of Hindi. The data analysis was done on DSP Sona-Graph whose sampling rate is 16 kHz. The parameter extraction was done on real time. The results from this analysis are summarized in the following sections.

5.4 Inherent F_0 of vowels

Each syllable nucleus (vowel) was associated with a specified average F_0 and is called inherent F_0 of the vowel. Within each test word further regular variations of F_0 were observed by changing the preceding or the following consonants, or the preceding or the following syllables (Lehiste & Peterson, 1961; Umeda, 1981; Ohde, 1984; Zawadsky & Gilbert, 1989). In the following section we discuss each of these issues separately.

There is a correlation between the height of the vowel and its inherent F_0 (Petersen & Barney, 1952; Lehiste & Peterson, 1959; Lehiste, 1970). If other

factors remain constant, high vowels (/i/ and /u/) exhibit high values of F_0 than low vowel (/a/). Physiologically this can be explained as follows. In the articulation of high vowels, the tongue is raised towards the roof of the mouth. The muscles constituting the tongue are attached to the superior part of the hyoid bone and some laryngeal muscles are attached to the inferior part. When tongue is raised, the larynx tends to be pulled upwards and laryngeal muscles are stretched (Lehiste, 1970). As mentioned earlier, the rise in the tension of laryngeal muscles results in the increase in F_0. The study shows that in Hindi the difference between high vowels and low vowel is about 20 to 30 Hz. In our experiments we noticed that the inherent F_0 of medium vowels (/e:/ and /o:/) are about 5 to 15 Hz from high vowel. If the quantity of the vowel is increasing without changing any other factor, then the inherent F_0 is also increasing. For example, the long vowel /a:/ has higher values of F_0 than the shorter counterpart /a/. It holds true for /i:/ and /u:/ also. In a disyllabic word the F_0 of the final vowel is greater than the F_0 of the initial vowel. This can be explained in terms of local fall-rise patterns. Fig.5.1. shows the inherent F_0 of each vowel for both word initial and final positions. In all cases the final vowel has greater values of F_0 than the initial vowel.

5.4.1 Effect of preceding consonant

The change in inherent F_0 of vowel with respect to the change in the preceding consonant was analyzed. The results are given in Appendix C19. Following are the conclusions made from this analysis.

For the analysis, preceding consonants were grouped based on the voicing nature and manner of articulation, and based on the place of articulation. While considering preceding consonants on the basis of voicing nature and manner of articulation, we are able to classify the changes in the values of F_0 into different consonant classes like unvoiced unaspirated stops (uvua), unvoiced aspirated stops (uva), voiced unaspirated stops (vua), voiced aspirated stops (va), nasals (nas), trills (tri), laterals (lat), semivowels (svow) and fricatives (fric). But we could not find any consistent change of F_0 value with respect to the place of

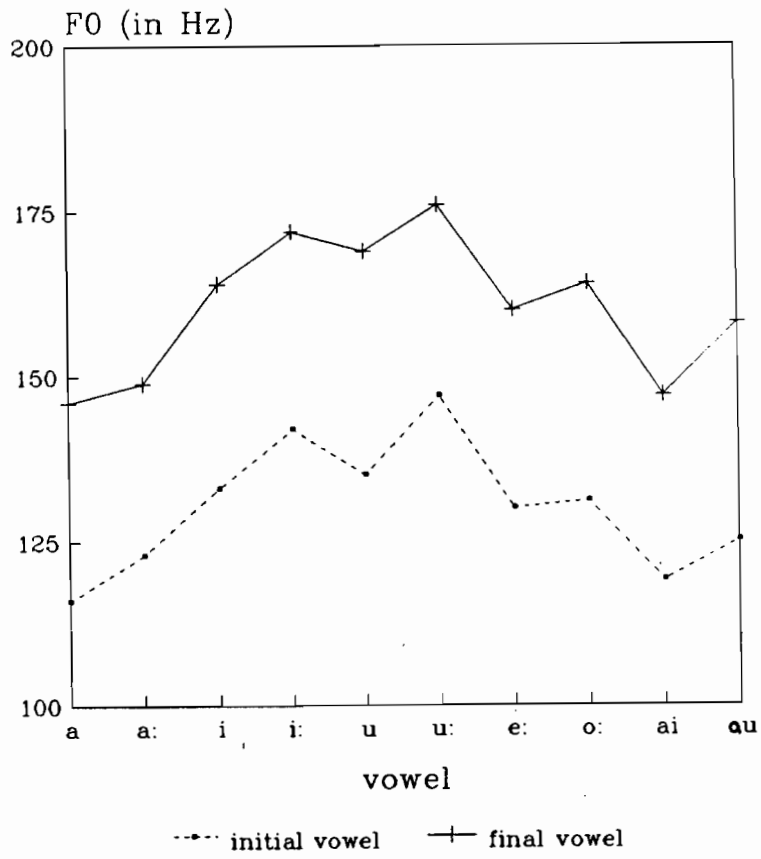


Fig.5.1. Inherent F_0 of vowels. Inherent F_0 of a vowel is proportional to the quality and the quantity of the vowel. High vowels exhibit higher F_0 than low vowel. Inherent F_0 of longer vowels are more compared with their shorter counter parts.

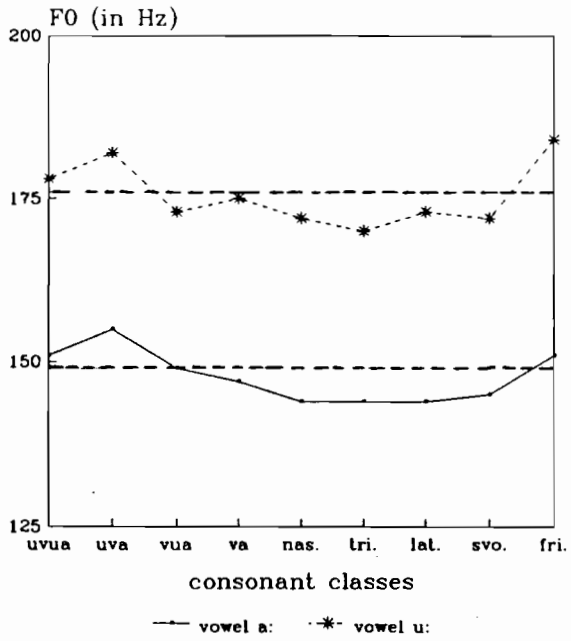
articulation of the preceding consonants. The different consonant classes in this category are bilabial (bil), dental (den), alveolar (alv), retroflex (ret), palatel (pal), velar (vel) and glottal (glo).

In general, higher values of F_0 occur after voiceless consonants and F_0 will be considerably low after voiced consonants. In our experiments we noticed that the maximum value of F_0 for a vowel occurs when it is preceded by voiceless fricative or voiceless stop, both aspirated and unaspirated. F_0 of the vowel is minimum if the preceding consonant is voiced, especially for nasals, trills and laterals. In the case of voiceless consonants the maximum F_0 gets shifted towards vowel onset position and in case of voiced consonants F_0 rises slowly and the peak occurs approximately in the middle of the syllable. Like voicing, aspiration of the preceding consonant also shows some effect on the inherent F_0 of the vowel. When the preceding consonant is aspirated, the F_0 of the vowel increases slightly (2 to 5 Hz). This was obvious in both voiced and voiceless cases with few exceptions. Due to the change in preceding consonant the inherent F_0 changes upto 15 Hz for different vowels.

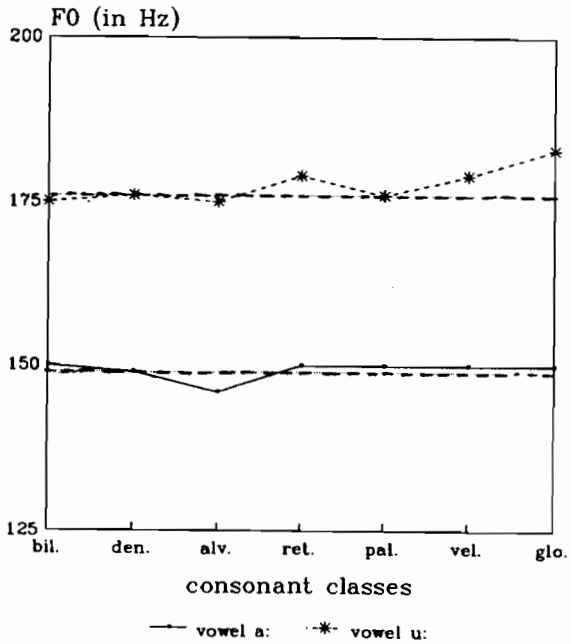
Fig.5.2 shows the influence of preceding consonant upon inherent F_0 of the vowels */a:/* and */u:/*. The inherent F_0 of these vowels are plotted against the different consonant classes based on voicing nature and manner of articulation (Fig.5.2a) and based on place of articulation (Fig.5.2b). The straight lines in both the cases show the average inherent F_0 of the corresponding vowel. From Fig.5.2a, it is seen that the change in inherent F_0 with respect to the voicing nature and manner of articulation of consonant classes is consistent for both the vowels.

5.4.2 Effect of following consonant

Like the preceding consonants, the inherent F_0 of the present vowel changes with respect to the changes in the following consonants. Appendix C20 shows the change in inherent F_0 with respect to the change of the following consonant in test words. For this analysis, we placed vowels at word initial position. Results of this analysis are concluded as follows.



(a)



(b)

Fig.5.2. Effect of the preceding consonant on inherent F_0 of the vowels */a:/* and */u:/*

(a) When consonants are classified based on the voicing nature and manner of articulation

(b) When consonants are classified based on the place of articulation

Straight line indicates the average F_0 of the corresponding vowel

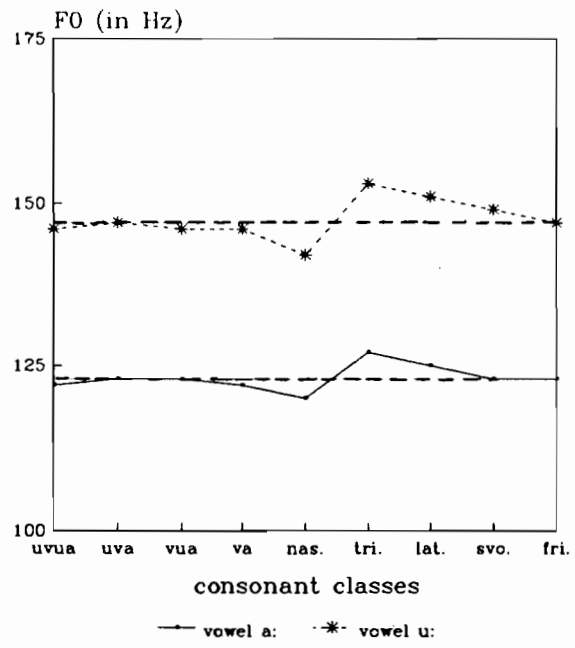
Following consonants are grouped and analyzed based on the voicing nature and manner of articulation, and on the place of articulation. Vowels did not show any consistent change in inherent F_0 when the following consonants are grouped based on the voicing nature and manner of articulation. But, there exists some changes with respect to the place of articulation of the following consonant. Inherent F_0 of a vowel appears to be more if the following consonant belongs to alveolar or palatel group. The inherent F_0 of the vowel is lesser than the average F_0 if the following consonant is dental or bilabial. Inherent F_0 of vowels change upto 10 Hz due to the changes in the following consonant.

Fig.5.3 shows the influence of the following consonants upon inherent F_0 of the vowels */a:/* and */u:/*. Inherent F_0 of the vowels are plotted against different consonant classes based on the voicing nature and manner of articulation (Fig.5.3a) and based on the place of articulation (Fig.5.3b). The straight lines show the average inherent F_0 of respective vowels. From the Fig.5.3b, it is seen that the changes in inherent F_0 of vowels are consistent in both the cases with respect to the change in the place of articulation of the following consonant.

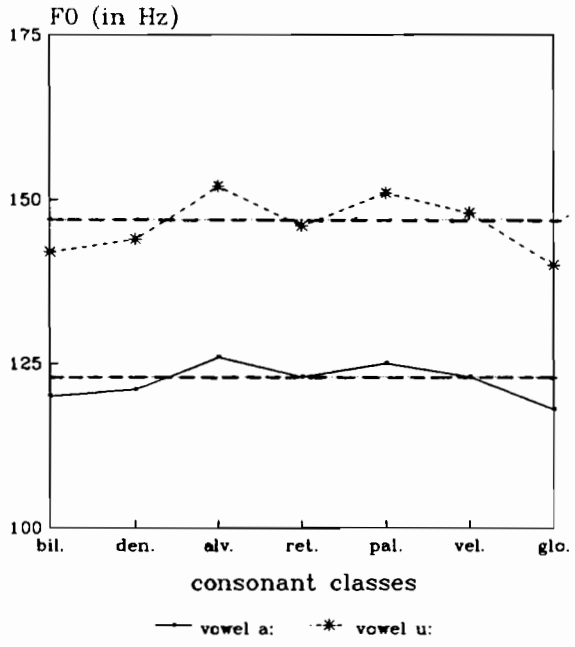
5.4.3 Effect of preceding syllable

If the nucleus of the preceding syllable changes, the inherent properties of present vowel also changes (Peterson, 1980). Appendix C21 shows the change in inherent F_0 of the vowel due to the change in the nucleus of the preceding syllable. Following are the conclusions made from this analysis.

Inherent F_0 of a vowel is affected by the quantity and quality of the previous syllable. F_0 of the present vowel changes with the height (quality) of the previous syllable nucleus. For example, in a disyllabic word, inherent F_0 of the final syllable will increase if the nucleus of the initial syllable changes from low vowel to high vowels. In magnitude, this change is upto 10 Hz. Like vowel height, duration of the previous syllable nucleus (quantity) also affects the inherent F_0 of the vowel. If all other things remain constant, then inherent F_0 of the vowel inversely changes with respect to the duration of the previous syllable. That is, inherent F_0 of the vowel is more if the previous syllable nucleus is shorter. This



(a)



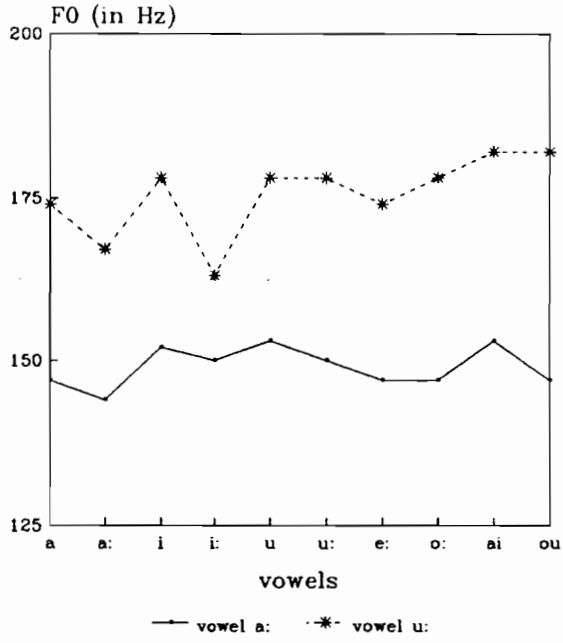
(b)

Fig.5.3. Effect of the following consonant on inherent F_0 of the vowels */a:/* and */u:/*

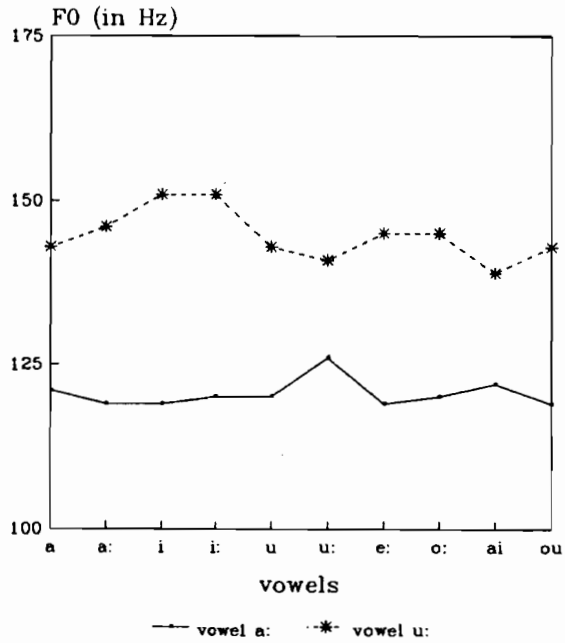
(a) When consonants are classified based on the voicing nature and manner of articulation

(b) When consonants are classified based on the place of articulation

Straight line indicates the average F_0 of the corresponding vowel



(a)



(b)

Fig.5.4. Effect of the adjacent syllables on inherent F_0 of the vowels */a:/* and */u:/*

(a) Changes in inherent F_0 due to the change in the nucleus of the preceding syllable

(b) Changes in inherent F_0 due to the change in the nucleus of the following syllable

has been observed for all occurrences of short vowels (*/a/, /i/* and */u/*) and their longer counterparts (*/a:/, /i:/* and */u:/*). Inherent F_0 of the vowel changes from 3 to 10 Hz due to the change in the duration of the previous syllable.

Fig.5.4a shows the effect of the preceding syllable on inherent F_0 of vowels */a:/* and */u:/*. The inherent F_0 is plotted against the nucleus of the previous syllable. The changes in inherent F_0 due to the change in the preceding syllable are consistent for all the vowels.

5.4.4 Effect of the following syllable

Following syllable did not show any consistent effect on the inherent F_0 of the present vowel. Appendix C22 shows the changes in inherent F_0 with respect to the change in the nucleus of the following syllable.

Fig.5.4b shows the effect of following syllable nucleus on inherent F_0 of the vowels */a:/* and */u:/*. There exists small changes with respect to the change of the following syllable nucleus. But these changes are not consistent for all the vowels and it is very difficult to capture by any general rule. Also magnitude of the change is very small compared with other effects. So these changes are ignored.

5.5 Summary

Influence of segmental properties on F_0 contours of speech sounds were discussed in this chapter. Each syllable nucleus is associated with a specific average F_0 and is called inherent F_0. Inherent F_0 of a syllable nucleus changes based on its phonetic properties and the properties of surrounding speech units. These properties were analyzed by controlled experiments in which nonsense disyllabic test words are embedded in a carrier sentence. This controlled study put severe limitations to the observations and it may not always match with the properties of F_0 contours of spontaneous speech. But generalising the properties of F_0 contour from spontaneous speech is very difficult.

Having discussed the global, local and segmental features of F_0 contour in Chapters 3 to 5, we discuss the applications of intonation knowledge on various speech systems, such as text-to-speech, speech-to-text and speaker recognition systems in the following chapters.

Chapter 6
SIGNIFICANCE OF INTONATION KNOWLEDGE FOR A TEXT-TO-SPEECH SYSTEM FOR HINDI

6.1 Introduction

Intonation knowledge occupies an important position in the speech signal since the prosodic information is predominantly determined by this parameter. Even though the prosodic knowledge by itself does not contribute to the speech information, it helps to organize the knowledge at segmental and at suprasegmental level to determine pitch, gain and duration of the sequence of basic units. Proper incorporation of intonation knowledge will enhance the naturalness and intelligibility of the synthetic speech significantly. In this chapter, we discuss the issues in the incorporation of intonation knowledge for a text-to-speech system for the Indian language, Hindi.

This chapter is organized as follows: Section 6.2 discusses a model for the incorporation of intonation knowledge. It includes a discussion on the design issues involved in the development of a text-to-speech system. Section 6.3 discusses the issues in the incorporation of intonation knowledge. Various issues involved in the representation and activation of intonation knowledge is discussed in this section. The quality of the synthetic speech output after the incorporation of intonation knowledge is evaluated in section 6.4.

6.2 The model: A text-to-speech system for Hindi

A text-to-speech system accepts an unrestricted text and converts it into speech waveforms. These systems typically use smaller stored units. They allow

utterances to built up from a finite set of small units and hence are able to generate speech corresponding to any input text. They also permit the modeling of a wide range of phonetic as well as linguistic details which are essential for the production of natural and intelligible synthetic speech (e.g., Umeda, 1976; Klatt, 1976; Hertz, Kadin & Karplus, 1985; Dochery & Shockey, 1988). Several successful unrestricted text-to-speech systems have been developed in languages like English (e.g., Klatt, 1987; Allen, 1976), French (O'Shaughnessy, 1984), Chinese (Lee, Tseng & Young, 1989), Japanese (Sato, 1984), etc..

We are developing a text-to-speech system for the Indian languages based on parameter concatenation model (Yegnanarayana, Murthy, Sundar, Alwar, Ramachandran, Madhukumar & Rajendran, 1990). Since speech has been modeled using parameters, voice characteristics can be manipulated and thus prosodic features can be incorporated by changing these parameters. This representation is highly flexible and needs much less storage compared to the waveform concatenation model.

The design of our text-to-speech system is modular. It enables us to make the changes in different modules. Later these modules were integrated to the rest of the system. Each of the modules added to the system was developed in parallel. The various modules available in our system include the knowledge sources related to coarticulation phenomena (Ramachandran, 1992), duration (Rajesh Kumar, 1990) and intonation.

Fig.6.1 shows the block diagram of our text-to-speech system. The input to the system is Hindi text stored in the form of ISCII (Indian Script Code for Information Interchange) codes. The preprocessor scans the string of ISCII codes to locate abbreviations, numbers, dates and special symbols and replace them by their expansions in spoken form. Basic units are extracted from the expanded text using a simple parser. For synthesizing speech, parameters of the basic units of input text are concatenated using coarticulation rules that operate across adjacent basic units of speech. The gain contour is smoothed at the boundaries between adjacent basic units. The pitch contour is modified to incorporate the intonation knowledge for the sentence being synthesized. The

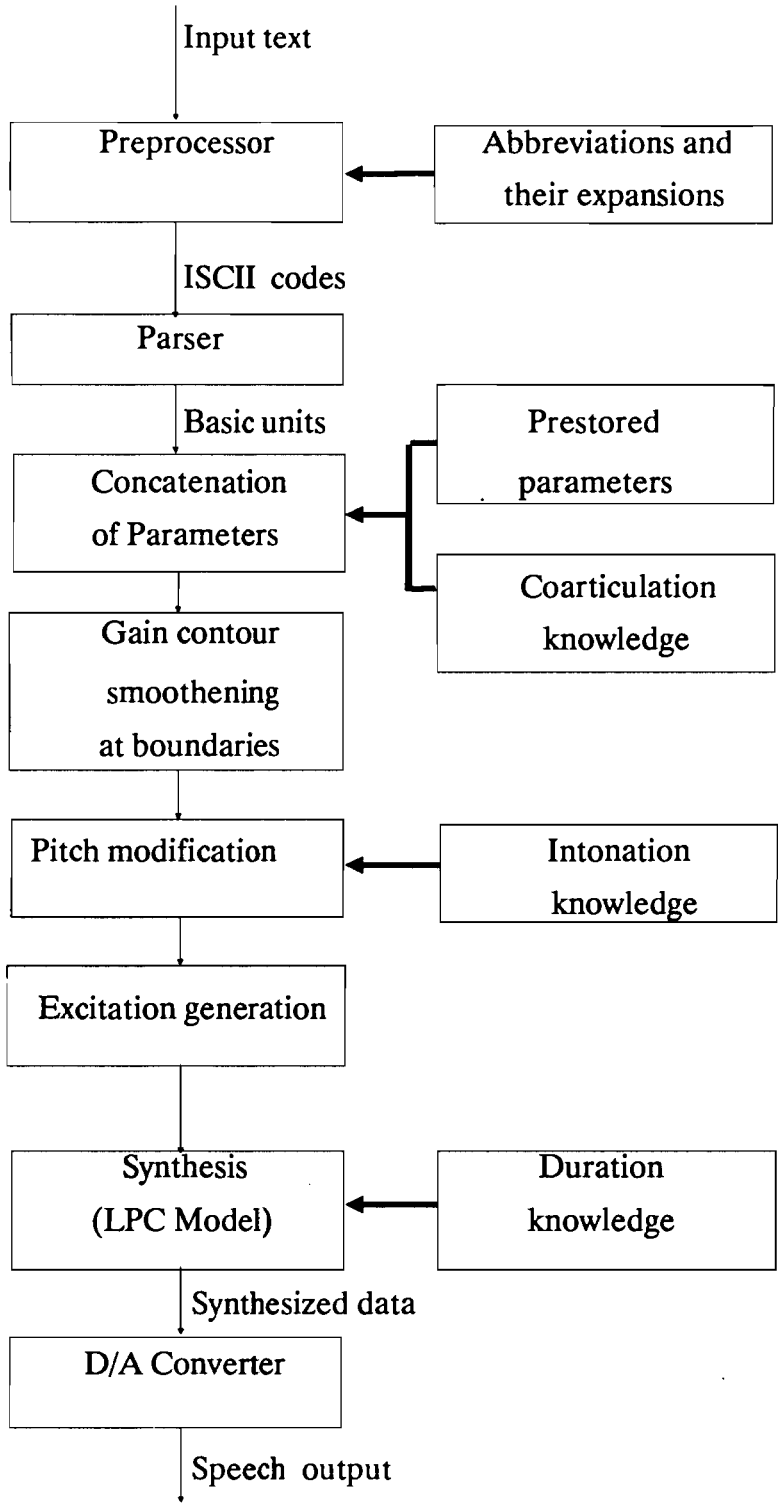


Fig.6.1. Block diagram of a text-to-speech system for Hindi

modified pitch and gain contours are used to generate an excitation signal. The excitation signal and the system parameters are used to generate the speech waveform.

Other than the incorporation of segmental and suprasegmental knowledge sources, following are the major issues involved in the design of our text-to-speech system: 1) choice of basic units, 2) collection of basic units and extraction of parameters, 3) preprocessor and parser and 4) synthesis of speech from the parameters of basic units. In the following sections we discuss each of these in detail.

6.2.1 Choice of basic units

The choice of basic units involves a trade-off between the size of memory needed to store all the units and the computation during synthesis. If the size of unit is large, the number of units in the language increases and hence the amount of computer memory needed to store them is also more. On the other hand if the size of unit is small, then the coarticulation effect among the adjacent units increases and which result in increased computation during synthesis (Macchi, Kahm & Streeter, 1987).

For Indian languages, the characters which are generally the orthographic representations of the speech sounds can be selected as a suitable choice for basic units. A character in an Indian language is close to a syllable and is more precise in definition. Character in Hindi represents a speech sound in the form of consonant (C) or vowel (V) or CV or CCV or CCCV. In characters, most of the coarticulation effects (all CV and CC transitions) are preserved. Also this can be extracted from the text by a simple parsing. Due to these reasons characters are chosen as basic units in the present implementation of our text-to-speech system.

We found out by experimental studies that the coarticulation effects by two adjacent consonants in a cluster is not significant. Therefore cluster characters can be generated from the constituent CV combinations and other consonants. For example, the cluster character */kya:/* can be generated by concatenating the

consonant /k/ and the CV combination /ya:/. This results in the reduction of the number of basic units (from 5000 to 400) and the storage requirement. Therefore the basic units in our text-to-speech system are: 1) isolated consonants (C); 2) isolated vowels (V) and 3) the consonant-vowel combinations (CV).

6.2.2 Collection of basic units and extraction of parameters

The basic units were extracted from the carrier words in isolation. The carrier words selected are of meaningless words to avoid the undesirable prosodic bias introduced subconsciously by the speaker. Also it allows us to quickly form a suitable carrier word to make the extraction of the basic units easier. The required basic unit is placed in the word medial position followed by a stop consonant with some exceptions (Rajesh Kumar, 1991).

To synthesize speech from a given text, our text-to-speech system uses following speech parameters: 1) LPCs, 2) formants, 3) pitch and 4) gain. LPCs and formants represent the vocal tract system and pitch and gain parameters correspond to the source information. In the following paragraphs we discuss the extraction of these parameters briefly.

Our text-to-speech system is based on linear prediction method. A set of 14 LPC parameters are used to model the vocal tract system. These are computed using autocorrelation method (Makhoul, 1975). The coarticulation effect is manifested in the speech wave mainly as transition pattern of formants (the resonant frequencies of vocal tract) (Ohman, 1966). A difficult signal processing problem is encountered in incorporating the coarticulation rules due to the incompatibility of the parameters used for basic units (LPCs) and for specification of the rules (formants). In order to solve this problem, all basic units are converted to a representable scheme in which the vowel regions are stored using formants and the consonant regions using LPCs. Formants are extracted using the properties of group delay functions (Yegnanarayana & Murthy, 1988).

We modify the intonation pattern by incorporating intonation knowledge obtained from the analysis of continuous speech in Hindi. Hence it is enough to

know that whether a frame of basic unit is voiced or unvoiced instead of the accurate pitch value. So voiced/unvoiced decision of the basic units are stored separately. Later, based on the intonation knowledge we modify the pitch contour of the utterances in voiced region. In the unvoiced region the pitch value is taken as zero.

We are using different methods for computing gain for consonants and vowels. Gain contour for each consonant frame is determined from the residual obtained by the autocorrelation method (Makhoul, 1975). In vowel portion, gain for each segment is computed as the sum of squared values of the signal. In order to solve the problems due to the incompatibility of gain computation, gain of each basic unit is pre-edited and stored. During the synthesis, after concatenating the parameters, gain contour is smoothed by the interpolation across the boundary of the adjacent basic units.

6.2.3 Preprocessor and parser

Preprocessor and parser are two preliminary modules in our text-to-speech system. The input and output of the preprocessor module is in ISCII code itself. The text is preprocessed to locate nonphonetic strings (such as numerals and abbreviations) which are replaced by their spoken form. For example, the abbreviation */dɑ̃:./* (Dr.) expands to its spoken form */da:ktar/* (Doctor) and the numeral 120.45 expands to */e:k sau bi:s dasamlav ca:r pa:nc/* (One hundred and twenty point four five). It also helps the intonation module to identify a particular word is numeral or not. The preprocessed ISCII codes are transferred to the parser.

Due to the phonetic nature of Indian languages, the parser module for our system is simpler than the languages like French and English where letter to sound rules and dictionary look-ups are used (O'Shaughnessy, 1984; Allen, Hunnicutt & Klatt, 1987). In this module sequences of ISCII codes are parsed to extract the sequence of basic units. The parser takes care of some language specific issues like word final vowel deletion (for short vowels in Hindi). The end of the sentence is identified by the presence of delimiters (e.g., bar (|), question

mark (?), etc.). These sequence of basic units are used for further processing to produce natural sounding, intelligible synthetic speech.

6.2.4 Synthesis of speech from the parameters of basic units

To synthesize speech from the parameters, we used linear predictive technique (Atal & Hanauer, 1971) in the consonant part and a cascade formant synthesizer (Klatt, 1980) in the vowel part of the basic units. By using this hybrid method for synthesis, it is possible to capture the coarticulation behavior of speech sounds as a set of transition patterns of formants in the vowel region.

The basic model for speech synthesis is given in Fig.6.2. It consists of an excitation signal and a time varying filter representing source and system parameters of speech signal, respectively. The excitation signal is periodic for voiced speech signals and a sequence of random numbers for unvoiced sounds. This excitation signal is fed to a time varying digital filter which models the vocal tract for generating speech. The time varying filter is represented by either LPCs or formants based on the type of basic units.

The choice of excitation signal affects the quality of synthetic speech significantly (Papamichalis, 1987). We used Fant's excitation model in our text-to-speech system. The Fant's model is supposed to resemble the actual glottal excitation (Fant, 1982; Childers, Ke Wu, Hicks and Yegnanarayana, 1989). Here, the energy is distributed evenly over the entire duration of the excitation. By varying opening and closing phase of excitation, it is possible to tune the quality of the output speech.

6.3 Incorporation of intonation knowledge in a text-to-speech system for Hindi

Our text-to-speech system has the flexibility to introduce prosodic variations in synthetic speech by varying the parameters of the basic units such as pitch, gain and duration. In this section, we examine the issues related to the incorporation of intonation knowledge in our text-to-speech system.

There are different stages required for incorporating intonation knowledge in a text-to-speech system. They are summarized as follows: (1) Input text has to

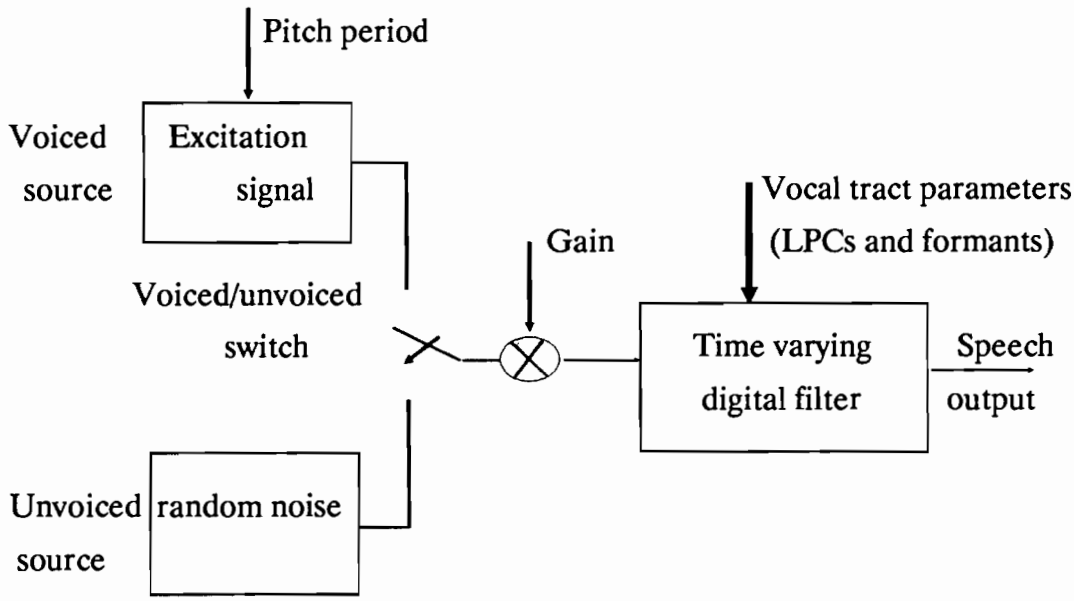


Fig.6.2. Block diagram for the basic speech production model. The vocal tract system is represented by LPCs and formants. Pitch period and gain are the source parameters. The voiced source of the speech is represented by an excitation signal and the unvoiced source is represented by random noise.

be parsed to find out the type of the sentence and the corresponding intonational properties as discussed in Chapters 3 and 4. (2) Text analysis has to be performed both at the word level and at the character level. Word level analysis decides the importance of each word in the sentence. Character analyzer determines the number of syllables in each word and it classifies the syllables based on the acoustic-phonetic behavior as discussed in Chapter 5. (3) Pitch accent patterns which includes the decision of valleys and peaks based on the pitch accent rules discussed in Chapter 3 have to be incorporated. (4) Effects of segmental properties on F_0 contour have to be incorporated. (5) Proper amount of pause has to be incorporated based on the various issues discussed in Chapter 4. (6) Intonation knowledge has to be represented using a suitable knowledge representation scheme in order to incorporate in a text-to-speech system. (7) Activation of intonation knowledge is achieved by means of a rule based inference engine with forward chain control strategy. Fig.6.3 places these seven issues in the overall scheme of our text-to-speech system. In the following sections, we discuss each of these issues in detail.

6.3.1 Intonation parser

Depending on the type of the sentence to synthesize, the global properties of F_0 contour is changing. For example, F_0 contour for simple declarative sentences show a declining tendency, while yes-no type interrogative sentences show a continuous rise. In complex sentences and in question-word type interrogative sentences F_0 contour exhibits a dual nature. The local attributes remain unchanged in all these cases. So in order to activate the intonation knowledge accurately, we need an intonation parser to determine the type of the sentence and the position of pattern change, if any. Intonation parser will take care of input sentences which have more than two words. If the number of words are less than three, then the pitch accent patterns are assigned to each word without considering the type of the sentence.

The detailed block diagram of intonation parser used by our text-to-speech system is given in Fig.6.4. Intonation parser consists of two stages of processing

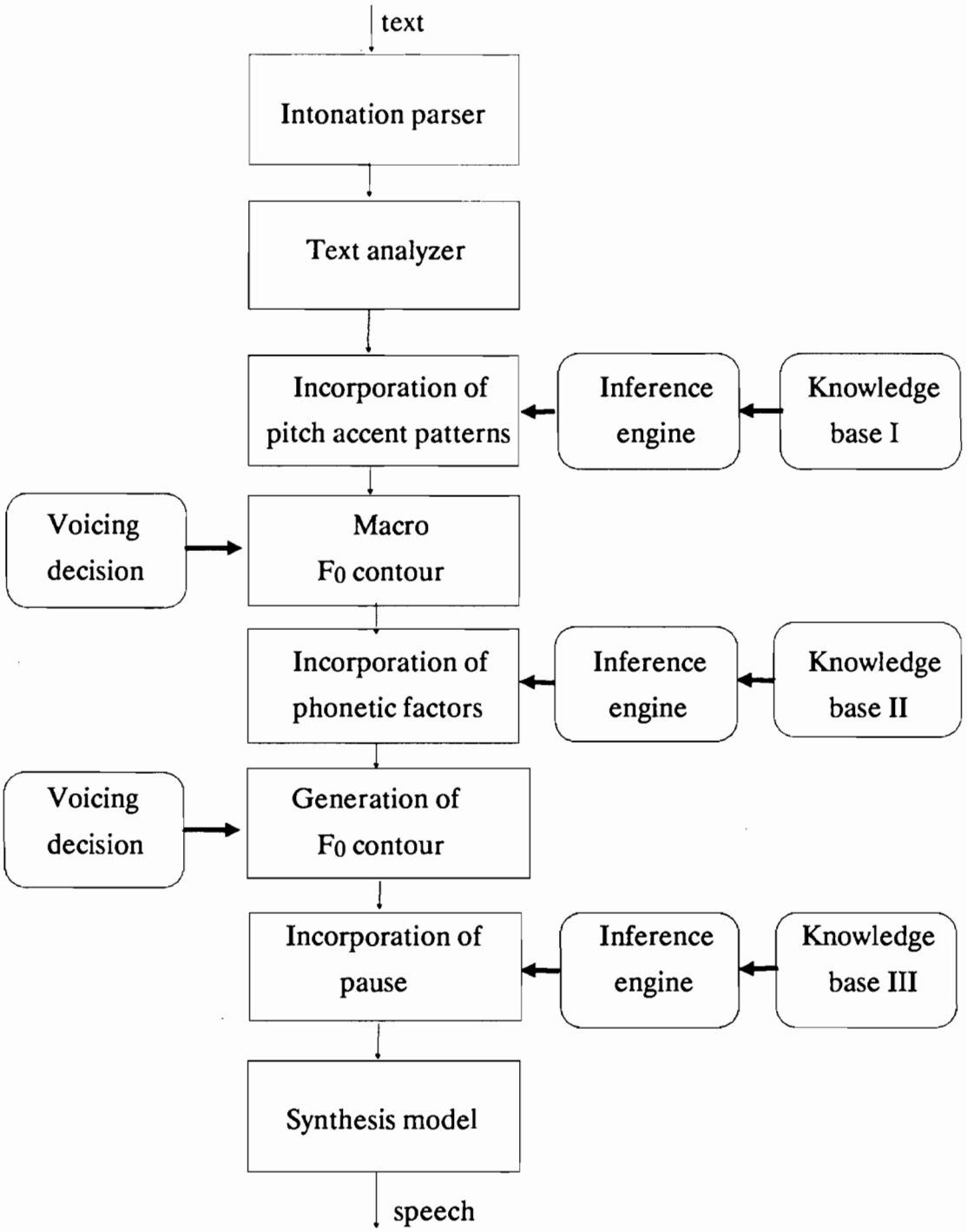


Fig.6.3. Incorporation of intonation knowledge in a text-to-speech system for Hindi.

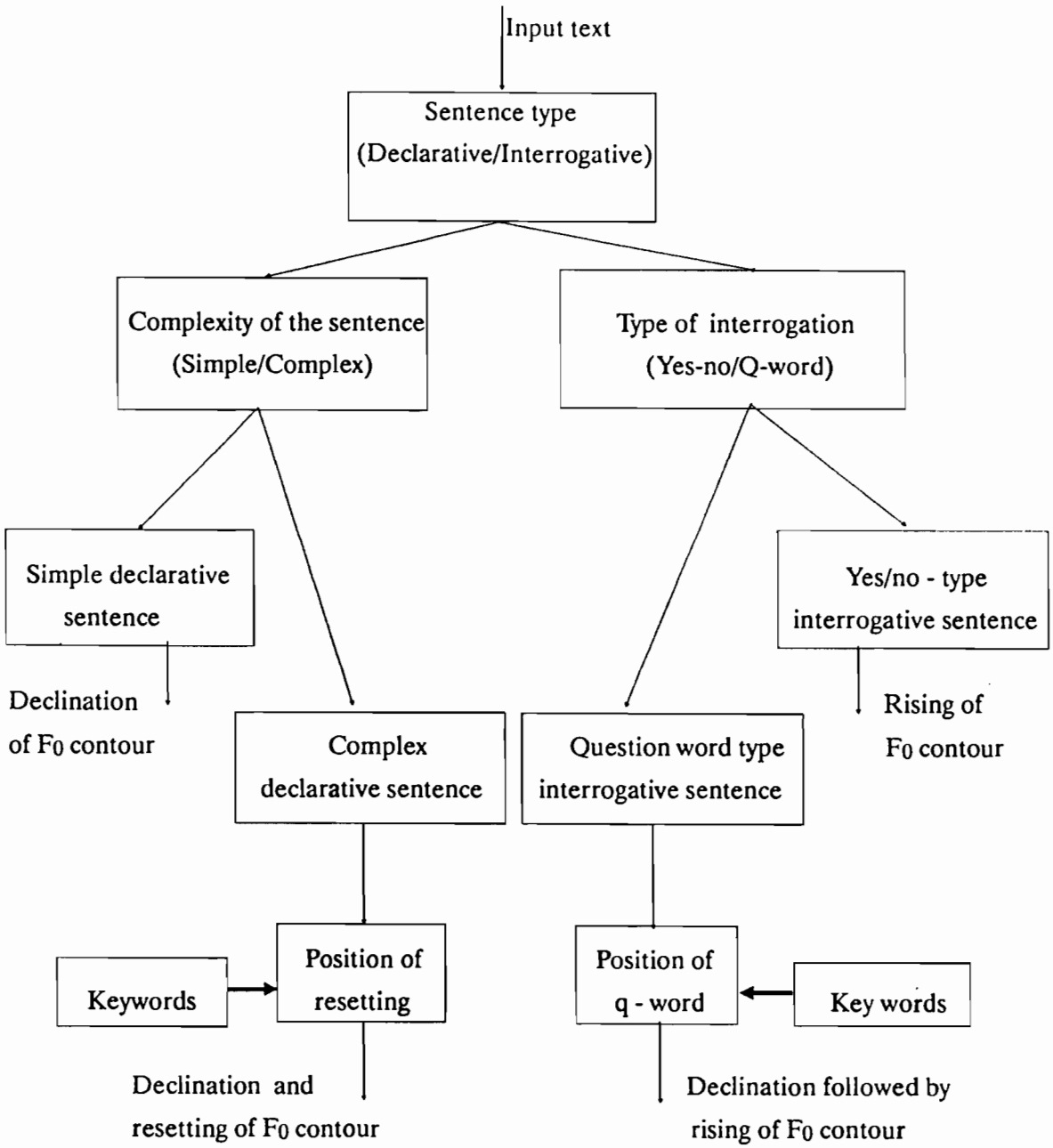


Fig.6.4. Block diagram of an intonation parser to determine the sentence types and to assign appropriate F_0 contour.

to determine the type of the sentence. Initially, the sentences are classified into declarative and interrogative sentences. The classification is based on the delimiters of the sentence. Since we did not consider other types of sentences for the time being, all types of sentences other than interrogative sentences are considered as declarative sentences. In the second stage, declarative sentence is divided into simple and complex sentences and interrogative sentences are divided into yes-no type and question-word type sentences based on key words. The global properties of F_0 contour assigned to each of these sentences are of different types as discussed in Chapters 3 and 4. Even though complex sentences are again divided into complex declarative and compound sentences, we have not considered these finer divisions to incorporate the intonation knowledge because in each of these cases, the value of F_0 resetting and pause between intonational phrases are almost equal (This is discussed in Section 4.3.2 in detail).

In complex sentences with more than two syntactic clauses, we input the text in such a way that each syntactic clause is separated by a comma. If the distance between two successive commas are less than three words, then the text between the commas are not considered as a syntactic clause. In all other cases, F_0 contour resets on the first word after the comma and the required pause is introduced before the word.

6.3.2 Text analyzer

The input text is analyzed to obtain the necessary information to enable the activation of intonation knowledge. This analysis can be divided into word level analysis and character level analysis. In the following sections we discuss each one separately.

6.3.2.1 Word analyzer

In order to incorporate the pitch accent patterns and other semantic aspects of F_0 contour a word analysis has to be done. Initially this analyzer divides the input words into monosyllabic function words and content words. Monosyllabic function word is again divided into three categories (independent, accented and unaccented function words) as discussed in Section 3.3.2.1. Independent

monosyllabic function words (function words which behaves like a content word) is included in the category of content words for assigning the pitch accent patterns. Function words of other two categories are conjoined with the preceding words accordingly. From the content words negation words and numerals are separated out because the presence of such words alter the F_0 contour as described in Section 4.3.3. The presence of numerals are detected by some special symbols communicated from the parser module. F_0 contour exhibits a dual nature for complex sentences and question-word type interrogative sentences. The position of pattern change is corresponding to some keywords as discussed in Chapters 3 and 4. The word analyzer detects corresponding keywords and marks the position for the change in F_0 contour which is used by the subsequent modules in the analysis. The block diagram of the word analyzer used by our text-to-speech system is given in Fig.6.5.

6.3.2.2 Character analyzer

In order to incorporate the inherent F_0 and other segmental prosodic variations, a character level analysis is necessary in the input text. In the present system, this analysis is performed to obtain the information about each of the following: (1) type of the basic units; (2) type of vowels and consonants in the basic units and (3) the number of characters in a word. Each of these are discussed in the following paragraphs.

Basic units in the given text can be of different types such as standalone consonant (C), standalone vowel (V), consonant-vowel combination (CV) or delimiters and other punctuation marks. The vowel and consonant components in the basic units are analyzed separately. Vowels are classified based on the quantity as well as quality. Details of this classification is given in Section 5.2.1. These classifications help us to incorporate the change in inherent F_0 of vowels with respect to their quantity as well as quality. Consonants are classified according to their manner and place of articulation. Details of this classification is discussed in Section 5.2.3. The changes in inherent F_0 due to the change in the preceding or the following consonant clauses can be incorporated using this

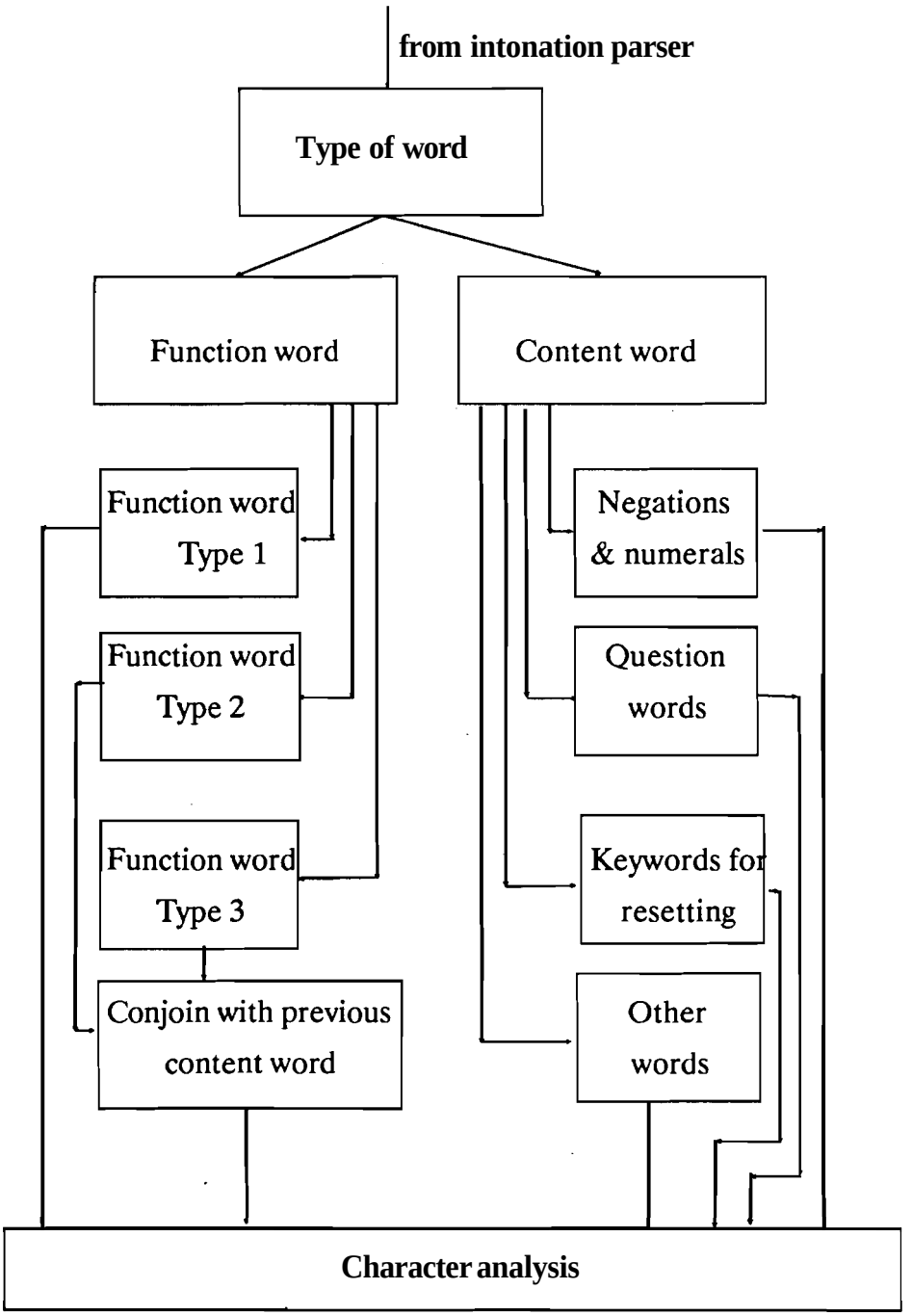


Fig.6.5. Block diagram for the word analyzer. Type 1 function word corresponds to the function words which have independent existance. Type 2 correponds to unaccented function words and Type 3 corresponds to accented function words.

classification.

Character analysis also finds the number of characters in a word. This will help us to determine whether the word is monosyllabic, disyllabic, trisyllabic, etc.. This information is passed to the next module to find out the valley and peak points based on pitch accent rules. Fig.6.6 shows the block diagram of the character analyzer used in our text-to-speech system.

Fig.6.7 shows the output from the intonation parser and the text analyzer modules for a simple declarative sentence */ajay ne: apni: tasvi:r de:khi:/* (Ajay saw his (own) picture). The intonation parser decides the type of the sentence. The text analysis routine categorizes words into function words and content words and find out the position of resetting. It also analyses the text into character level and classify them based on various phonetic properties of consonants and vowels.

6 3 3 Incorporation of pitch accent rules

Valleys and peaks for each prosodic word are decided based on pitch accent rules in Hindi (discussed in Section 3.3.2.1). Monosyllabic words contain a valley followed by a peak, both on the same syllable. Disyllabic, trisyllabic and monomorphemic tetrasyllabic words have valley on the initial syllable and peak on'the final syllable. Valley-peak assignment of bimorphemic tetrasyllabic and pentasyllabic words are not given in our system. For that we need a morphemic analyzer which was not taken as a part of this research study. Some content words like negation words and numerals can alter the general F_0 contour to some extent. So these words are taken separately and their valleys and peaks are assigned accordingly. A suitable knowledge representation scheme is used for representing these knowledge (discussed in Section 6.3.6) and is activated using an inference engine (discussed in Section 6.3.7).

After the incorporation of valleys and peaks, we have to model a macro F_0 contour for the incorporation of segmental properties of Fo contour. In order to generate a macro F_0 contour, we joined successive valleys and peaks using straight lines in the voiced portions of the corresponding basic units. For all voiced frames, the F_0 value is assumed as the value given by the straight line

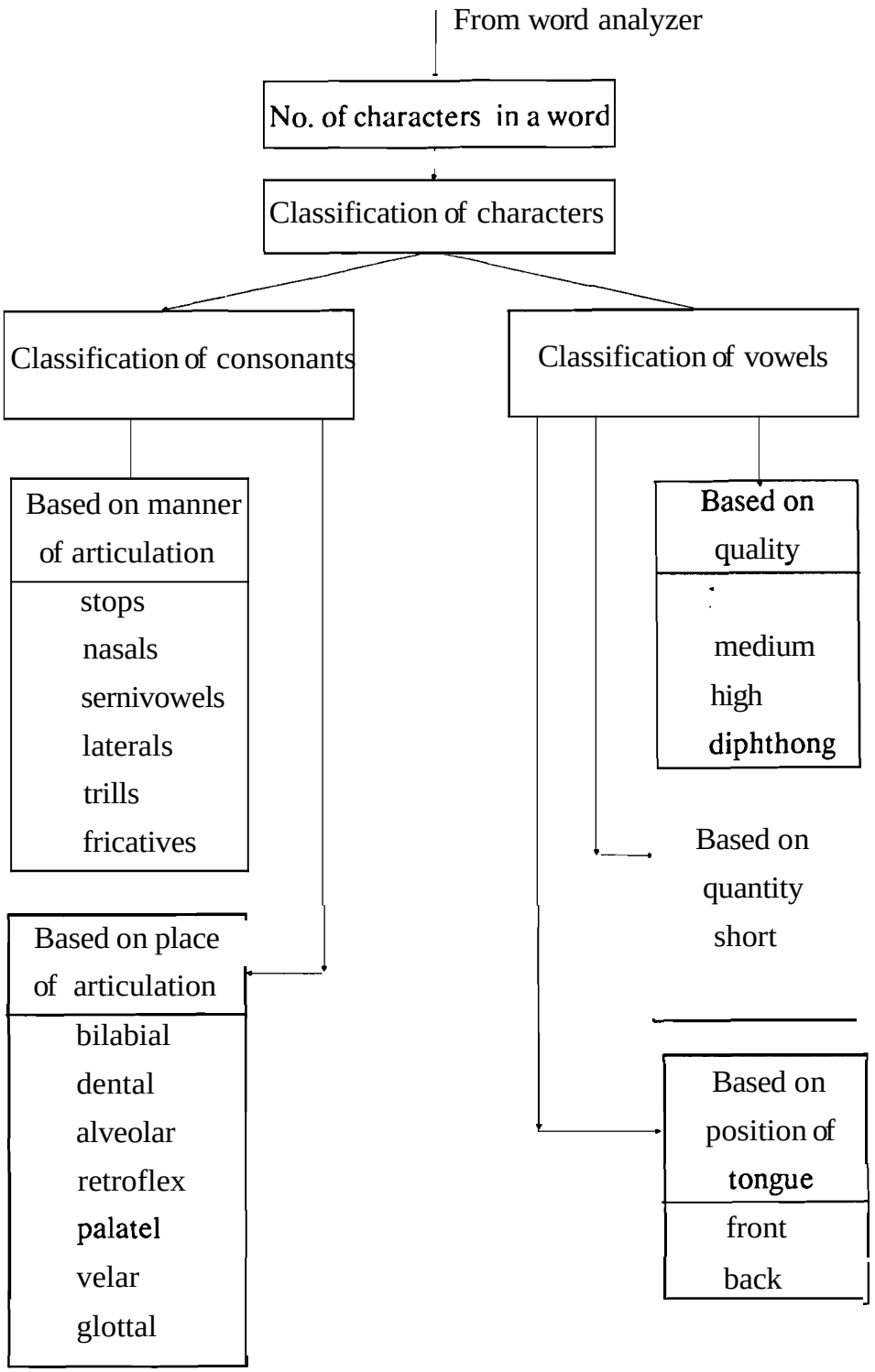


Fig.6.6. Block diagram of the character analyzer used in our text-to-speech system

Sentence	*a j a y*	*ne:*	*a p n i:*	*t a s v i: r*	*d e: kh i:*
Intonation parser	simple declarative sentence				
Word analyser	CW	FW2	CW	CW	CW
Character analyser					
No of characters/word	3	1	3	4	2
Consonant classificaton					
(i) manner of articulation	/j/ palate1	/n/ dental	/p/ bilabial	/t/ dental	/d/ dental
	/y/ palate1		/n/ dental	/s/ alveolar	/kh/ velar
				/v/ dental	
				/r/ retroflex	
(ii) place of articulation	/j/ VUA	/n/ nasal	/p/ UVUA	/t/ UVUA	/d/ VUA
	/y/ s.vowel		/n/ nasal	/s/ fricat.	/kh/ UVA
				/v/ s.vowel	
				/r/ trill	
Vowel classification					
(i) quality of vowel	/a/ low	/e:/ medium	/a/ low	/a/ low	/e:/ medium
	/a/ low		/i:/ high	/i:/ high	/i:/ high
(ii) quantity of vowel	/a/ short	/e:/ long	/a/ short	/a/ short	/e:/ long
	/a/ short		/i:/ long	/i:/ long	/i:/ long
(iii) position of tongue	/a/ back	/e:/ front	/a/ back	/a/ back	/e:/ front
	/a/ back		/i:/ front	/i:/ front	/i:/ front

Fig.6.7. Output of the text processing modules in our text-to-speech system. CW indicates content word and FW2 indicates unaccented function word. UVUA, UVA and VUA indicate unvoiced unaspirated stop, unvoiced aspirated stop and voiced unaspirated stop, respectively.

joined by the nearest valley and peak.

63.4 Incorporation of inherent F_0

Character level features of F_0 contours are incorporated in this module. The Fo at the mid point of each syllable nucleus obtained from the macro analysis are considered as the base inherent F_0. This F_0 value is modified based on the phonetic properties of vowels and the surrounding speech units. For example, when an voiceless consonant occurs in an accented syllable, the peak of the F_0 contour is slightly shifted towards the vowel onset position. Also, the F_0 at that position is about 5 to 15 Hz higher than the F_0 at the middle of the vowel. Knowledge obtained from the microprosodic analysis given in Chapter 5 are coded into a suitable form at this stage. This knowledge is activated by using an inference engine. However, if the change in F_0 contour is very small (less than 1% of the base inherent Fo), then such rules are ignored.

After the incorporation of the knowledge on inherent F_0, the final F_0 contour has to be smoothed. This smoothing can be done by spline curves (Rogers & Adams, 1989). The smoothed F_0 contour is used for synthesis.

6.3.5 Incorporation of pause

The final part of the incorporation of intonation knowledge is the incorporation of pause between syntactic clauses and between words. Pauses are determined by the factors discussed in Section 4.4, such as the number of syntactic clauses in a sentence, phonetic factors of words, etc.. For example, if the number of syntactic clauses are more, then the pause between them is less. All these rules are coded and represented in a knowledge base and activated accordingly. During a pause, all source and system parameters are set to zero. Insertion of proper pause will enhance the quality of synthetic speech to a large extent.

63.6 Representation of intonation knowledge

The intonation knowledge obtained from the analysis of natural speech has to be coded in a suitable form in order to incorporate in a text-to-speech system.

Our system is based on production system approach. Production systems are currently the most common knowledge representation techniques used in expert systems (Hayes-Roth, Waterman & Lenat, 1983; Rich, 1983; Browmston, Farrel, Kant & Martin, 1986). The knowledge is represented using IF-THEN rules. Each rule in the knowledge base is an independent fragment of knowledge and does not relay on the correctness of other rules. This facilitates successive updating because the rules are independent of each other and the order of declaration of rules are not important. Besides, rules in the production system provides an easy way to give an explanation for the intermediate decisions taken. The format of a rule is as follows:

IF <number of antecedents>
antecedent 1
antecedent 2

THEN <number of consequents>
consequent 1
consequent 2

For example, a rule could be as follows:

IF <2>
word is ***at sentence beginning***
word is disyllabic
THEN <2>
select the middle of the second vowel
fix the pitch accent-value as 180 Hz

The above rule states that if a disyllabic word occurs at the beginning of input text, then the pitch accent value at the midpoint of the second syllable nucleus (vowel) is 180 Hz.

63.7 Activation of intonation knowledge

Since we represented intonation knowledge as rules, the activation of the knowledge is achieved by means of a rule based inference engine (rule interpreter). Here the rules are classified into three: 1) rules related to pitch accent rules; 2) segmental prosodic variations of F_0 contour and 3) rules for the pause insertion. In our system, each of these rules is represented by separate knowledge bases.

For activating all this knowledge bases we used an inference engine of forward chaining control strategy. It is applied for all conditions specified in the knowledge base. From the knowledge base the rules are read into an array of records. The fields of this record contains total number of antecedents for each rule, list of antecedents, total number of consequents for each rule, list of consequents and state. The state field is a control flag and is used to mark a rule once it has been fired. Thus it prevents the same rule being fired again, and prevent the inference engine from entering into an infinite loop.

After the application of the rules, the speech is synthesized using the modified pitch contour. Quality of the speech output improved significantly. Improvement in the quality of synthetic speech after the incorporation of intonation knowledge is discussed in the next section.

6.4 Evaluation of the quality of synthetic speech

It is necessary to evaluate and compare the performance of each rule added to the system on the resulting quality of speech. Quality of synthetic speech is usually referred to the total auditory impression of the listener experiences upon hearing the speech from the system. The listener impression is influenced by the various constraints such as familiarity to the language, the inherent limitations of the human information system, the experience and training of human listener, the linguistic structure of the message set and the structure and quality of the speech signal (Pisoni, Nusbaum & Green, 1985; Childers & Ke Wu, 1990; Van Bezooijen & Pols, 1990; Monaghan & Ladd; 1990).

There is no well formed method for assessing the performance of synthetic

speech quality. Assessment of the perceptual response by the human listener investigate the transmission of linguistic information from the speech signal and address specific questions such as how accurately synthetic characters and words are recognized, how well the meaning of synthetic utterance is understood and how easy it is to perceive and understand synthetic speech. When we are considering intonation knowledge in particular, the perceptual questions arising are where the listener recognizes the properties of F_0 contours in Hindi. That is, declination/rising tendency, local fall-rise patterns and the beginning of a new structural domain if the system resets its F_0 declination. Following are the discussion of some perceptual experiments done by us.

In order to test the improvement in perceptual quality, we synthesize several sentences of different classes by using 1) waveform concatenation model, 2) parameter concatenation model and 3) parameter concatenation model with the addition of intonation knowledge. In waveform concatenation model, speech is retrieved from prestored digitized speech data corresponding to the basic units in the text and concatenate them in proper sequence to produce speech. In parameter concatenation model, the basic units are coded using parameters and the speech (corresponding to a given text) is synthesized from these parameters. The evaluation of the quality of synthetic speech was done based on the subjective measurements of native and non-native speakers of Hindi.

From our experiments, we found that the speech obtained from the parameter concatenation model (Type2) is better than the waveform concatenation model (Typel). There are abrupt discontinuities in the synthetic speech obtained from Type1 at the boundaries of the basic units. These are removed in Type2 as concatenation is now done at the parameter level resulting in a reasonably smooth transition between boundaries. But there are various distortions in Type2 due to the lack of proper prosodic knowledge in synthetic speech. The significance of intonation knowledge can be perceived in parameter concatenation model after incorporating intonation knowledge (Type3). Here, the intonation knowledge is coded into a suitable format and is activated as discussed in previous section. The quality has improved significantly. Depending

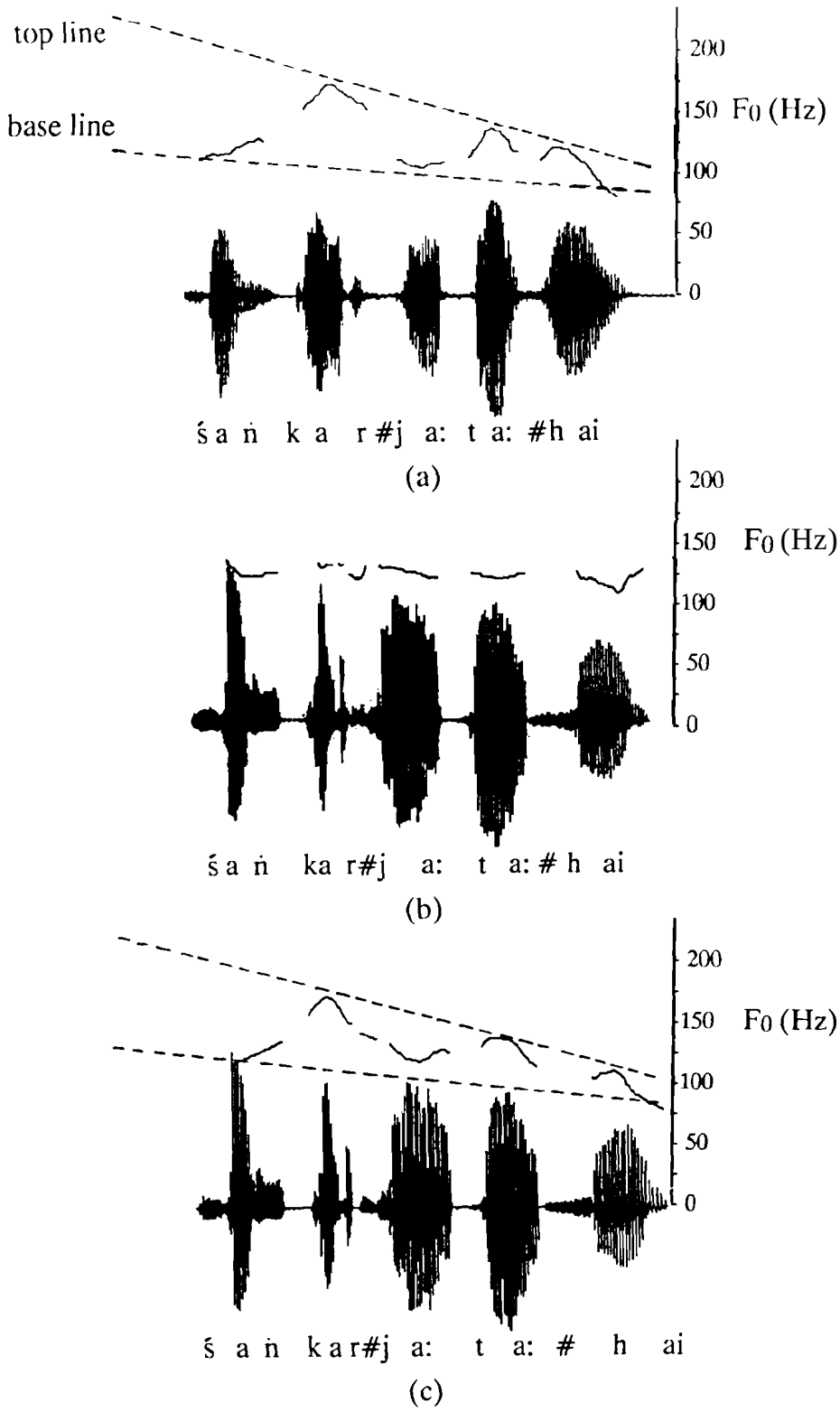


Fig.6.8. Speech waveform and Fo contour for a simple declarative sentence */śaṅkar ja:ta: hail* (Shankar goes)

(a) Natural speech signal

(b) Synthesized speech signal without applying intonation knowledge

(c) Synthesized signal after applying intonation knowledge

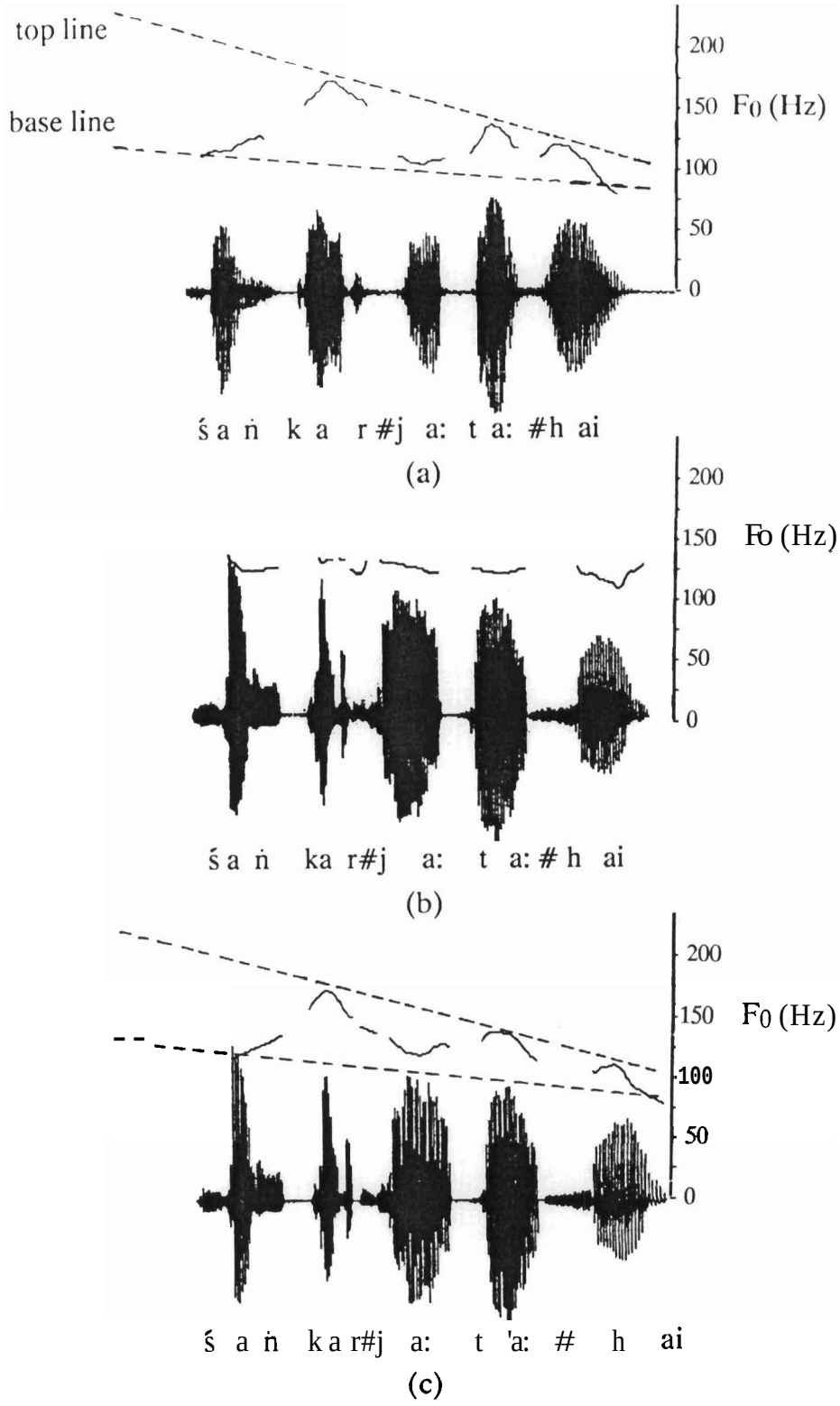
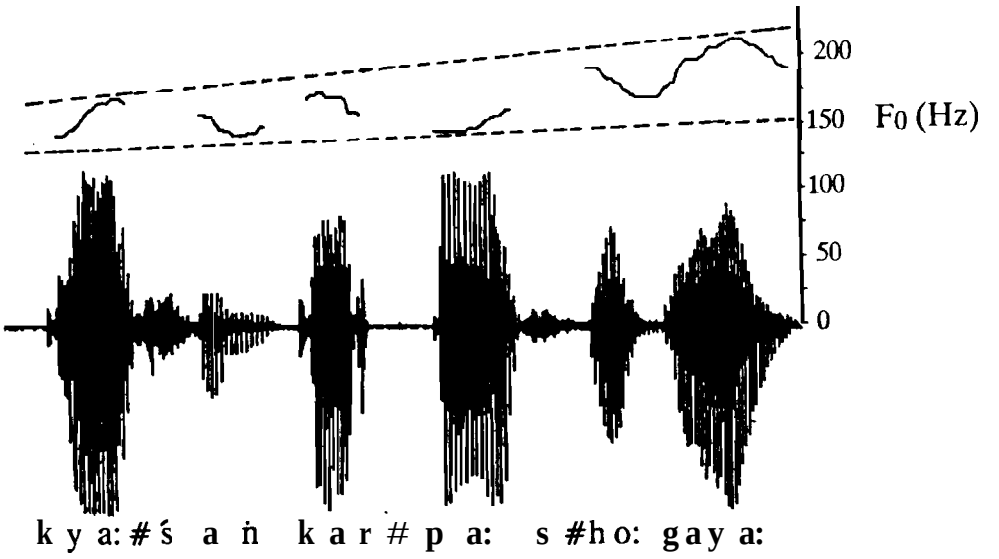


Fig.6.8. Speech waveform and Fo contour for a simple declarative sentence */śańkar ja:ta:* hail (Shankar goes)

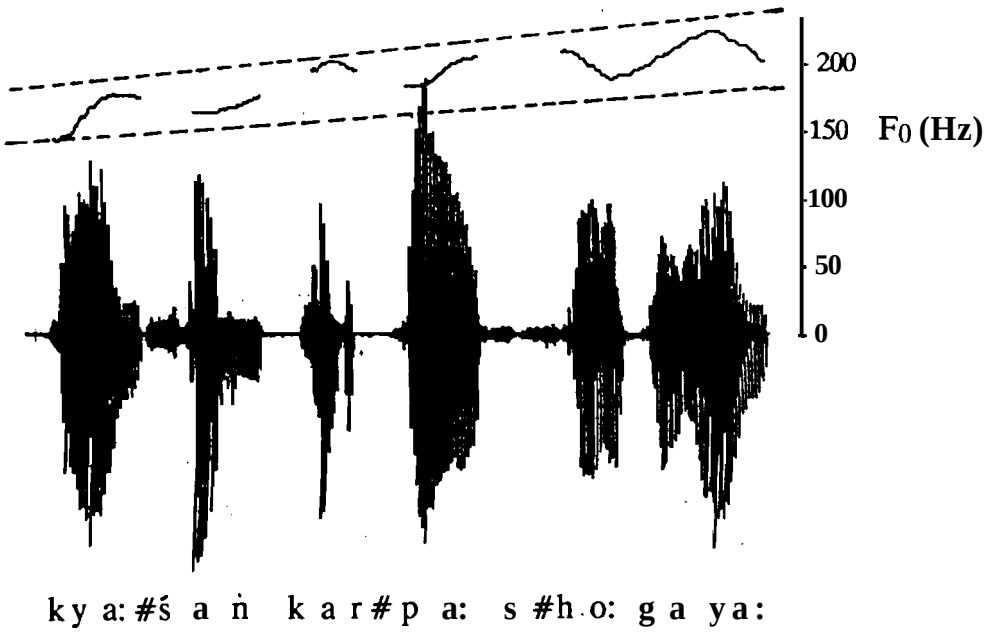
(a) Natural speech signal

(b) Synthesized speech signal without applying intonation knowledge

(c) Synthesized signal after applying intonation knowledge



(a)

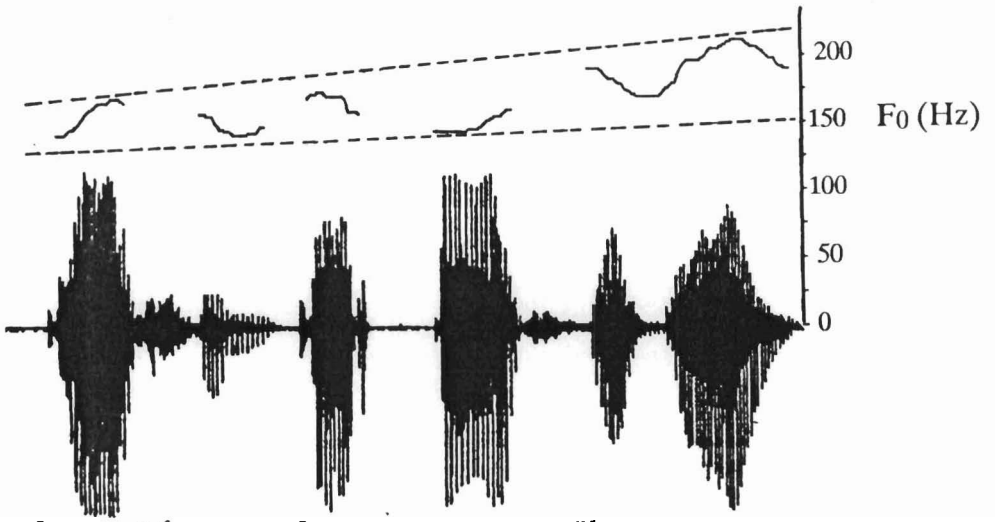


(b)

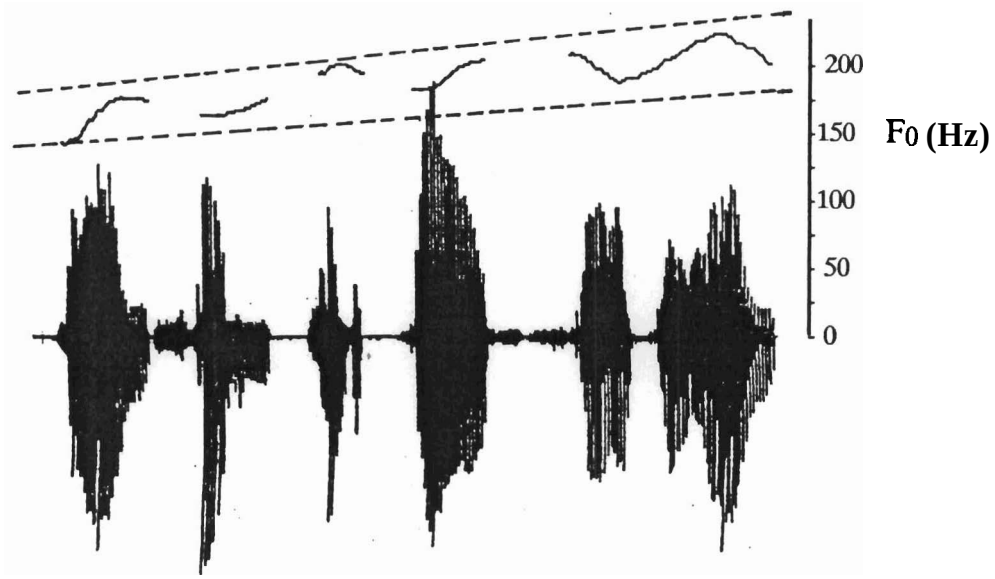
Fig.6.9. Speech waveform and Fo contour for an yes/no type interrogative sentence */kya: śaṅkar pa:s hogaya:?/* **(Has Shankar passed?)**

(a) Natural utterance

(b) Synthesized speech



(a)

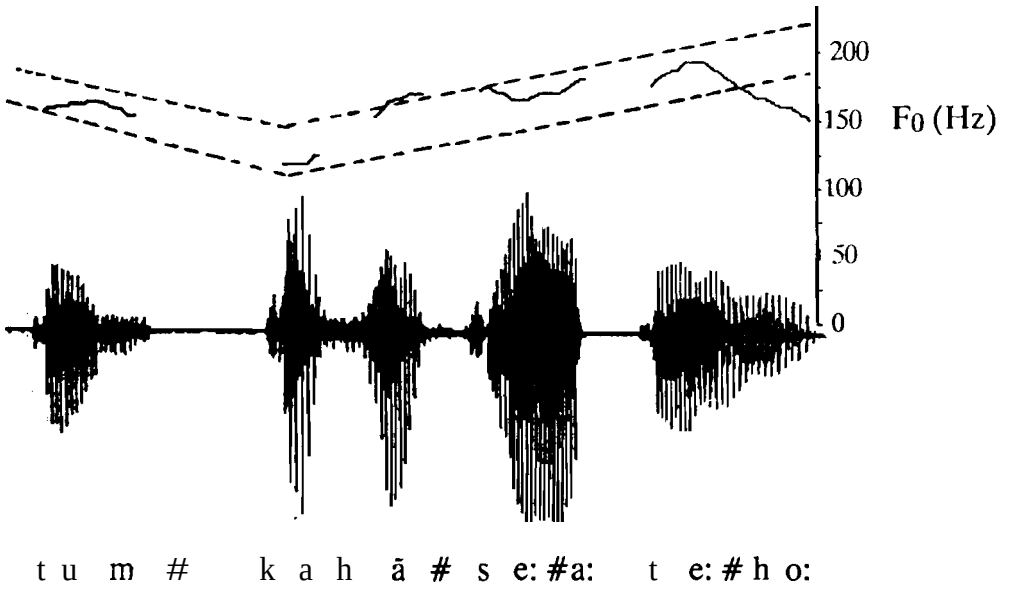


(b)

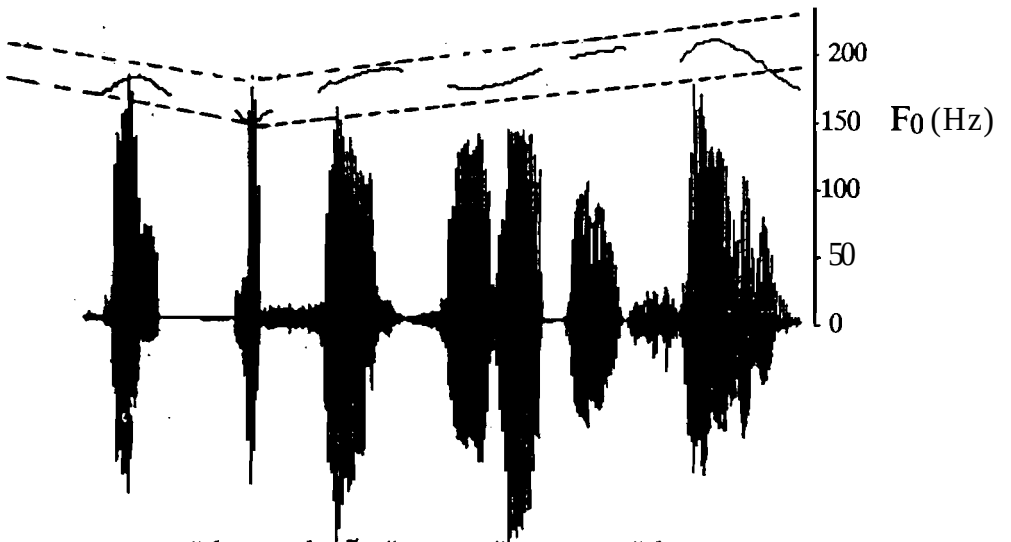
Fig.6.9. Speech waveform and F_0 contour for an yes/no type interrogative sentence /*kya: śaṅkar p a s hogaya:?*/ (Has Shankar passed?)

(a) Natural utterance

(b) Synthesized speech



(a)

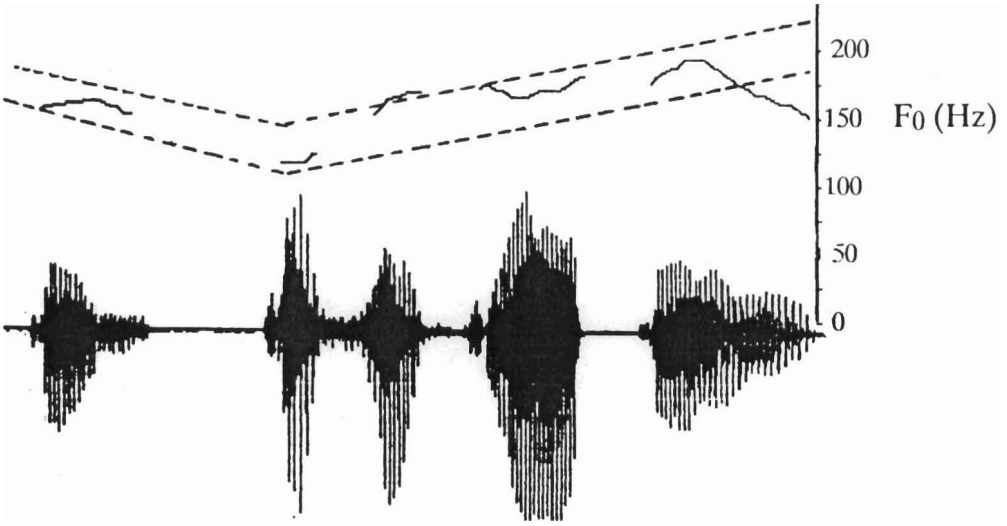


(b)

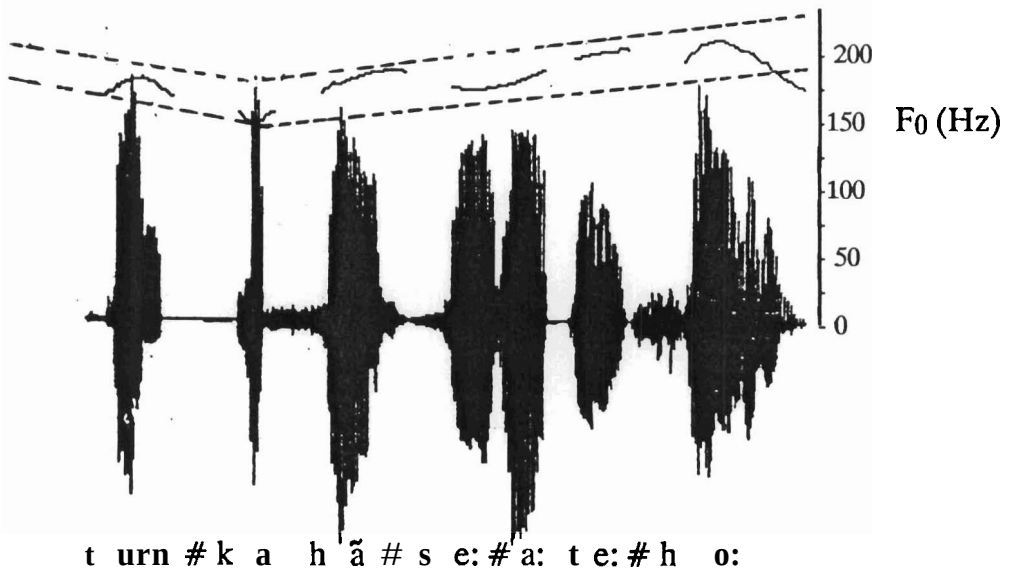
Fig.6.10. Speech waveform and F_0 contour for a question-word type interrogative sentence */tum kahã: se: a:te: ho:?/* (Where did you come from?)

(a) Natural utterance

(b) Synthesized speech



(a)

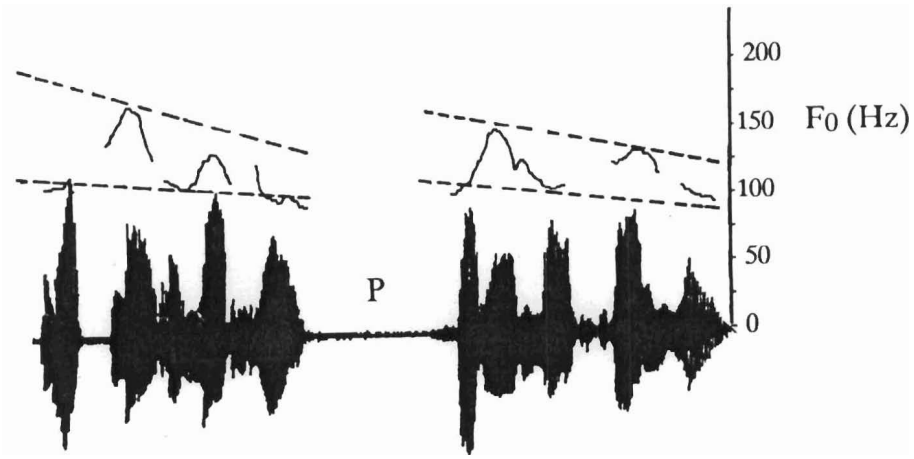


(b)

Fig.6.10. Speech waveform and F_0 contour for a question-word type interrogative sentence *[tum kahã: se: a:te: ho:?]* (Where did you come from?)

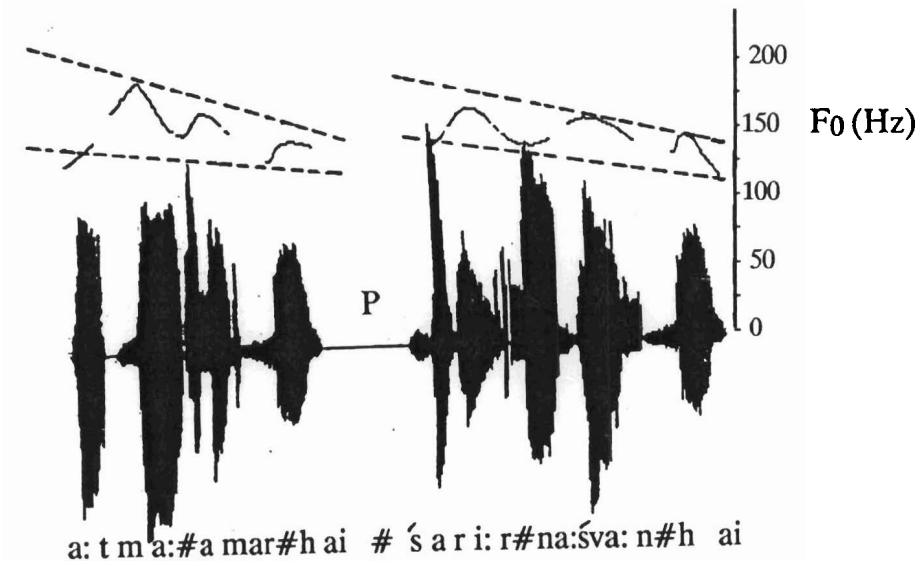
(a) Natural utterance

(b) **Synthesized speech**



a: t ma:#amar#h ai # ś a r i: r#na: śva:n# h ai

(a)



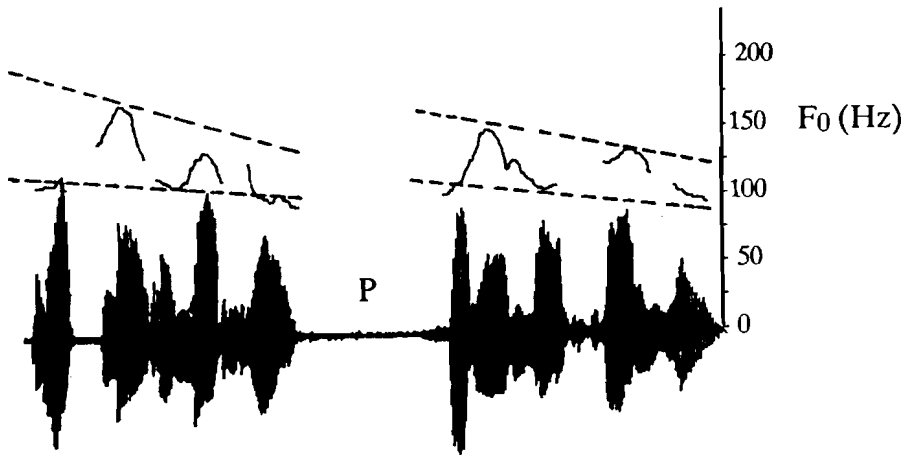
a: t m a:#a mar#h ai # ś a r i: r#na:śva: n#h ai

(b)

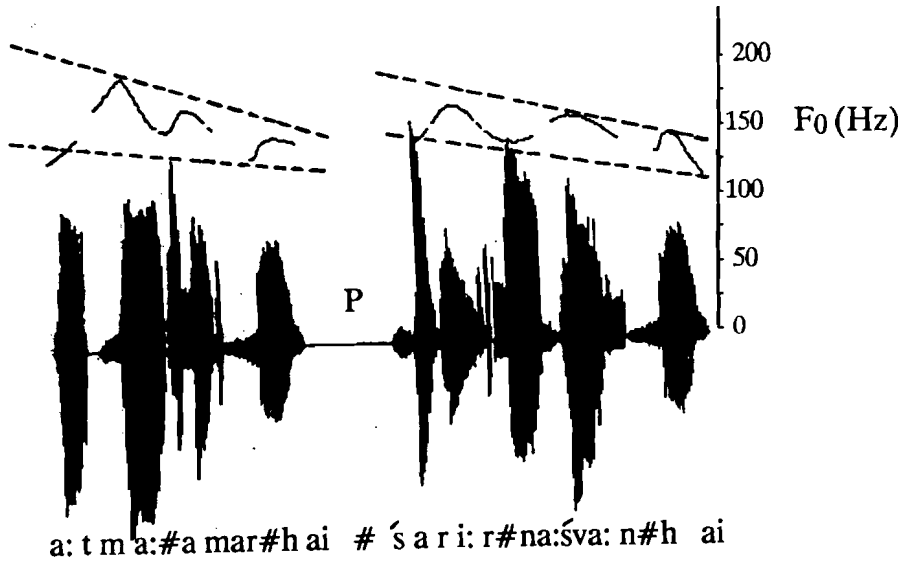
Fig.6.11. Speech waveform and F_0 contour for a complex declarative sentence */a:tma: amar hai śari:r na:śva:n hail* **(Soul is immortal, body is mortal)**

(a) Natural utterance

(b) Synthesized speech



(a)



(b)

Fig.6.11. **Speech waveform and F_0 contour for a complex declarative sentence** */a:tma: amar hai śari:r na:śva:n hail* **(Soul is immortal, body is mortal)**

(a) Natural utterance

(b) Synthesized speech

on the outcome of the performance of the system, we can modify the system further to attain the ultimate goal - to develop a text-to-speech system which produces the natural quality speech output.

Fig.8a shows the natural utterance and F_0 contour for a simple declarative sentence */śaṅkar ja:ta: hai/* (Shankar goes). Fig.8b shows the synthetic speech and corresponding F_0 contour for the sentence shown in Fig.8a. After activating the intonation knowledge, the F_0 contour gets modified as shown in Fig.8c. Similarly, Figs.9, 10 and 11 show the natural and synthetic speech signals and corresponding F_0 contours for yes-no type interrogative sentence, question-word type interrogative sentence and complex declarative sentence, respectively. In all these cases intelligibility and naturalness of synthetic speech are improved considerably by the addition of intonation knowledge.

6.5 Summary

This chapter addressed the issues in the incorporation of intonation knowledge in a text-to-speech system for Hindi and the evaluation of the performance after the incorporation of intonation knowledge. The text-to-speech system to which intonation knowledge was incorporated is based on parameter concatenation model. The system consists of two parts: (1) the analysis of input text and (2) synthesis of speech from the basic units together with modules for various knowledge sources such as coarticulation, duration and intonation. The intonation knowledge was incorporated into the text-to-speech system by changing the pitch contour of the sequences of basic units based on the rules related to the type and structure of input sentences, fall-rise patterns, inherent F_0 and pause. An intonation parser was developed to find out the type of the sentence and the appropriate intonation pattern. The required intonation knowledge is represented in production system format and activated using an inference engine of forward chain control strategy. The quality of speech output was evaluated. The naturalness and the intelligibility of synthetic speech improved significantly after the incorporation of intonation knowledge.

Chapter 7
APPLICATIONS OF INTONATION KNOWLEDGE FOR A SPEECH-TO-TEXT SYSTEM FOR HINDI

7.1 Introduction

In a speech-to-text system one has to use the acoustic parameters of continuous speech as well as the linguistic features of the language to transform a speech waveform to a set of symbols. The various issues involved in the development of a speech-to-text system are: (1) hypothesization of word boundaries from the continuous speech; (2) segmentation of the utterance into small speech segments of basic units; (3) analysis of speech segments based on their acoustic-phonetic properties; (4) lexical analysis to correct the output of acoustic-phonetic analyzer and the word boundary hypothesization and (5) syntactic and semantic analysis to correct the errors in lexical analysis and to produce a meaningful text from the given utterance. Fig.7.1 places all these issues in the overall scheme of a speech-to-text system. Of these, this chapter discusses the use of intonation knowledge for hypothesizing word boundaries from continuous speech.

A text without word boundaries is difficult to read. The readability of the transformed text improves significantly even if it contains a few word boundaries. Continuous speech recognition systems make use of many knowledge sources for placing word boundaries besides lexical matching. Some of these knowledge sources are phonotactic constraints (Harrington, Johnson & Cooper, 1987), durational clues (Wightman & Ostendorf, 1991), language clues (Ramana Rao & Yegnanarayana, 1991), syntactic constraints, etc.. Use of these knowledge sources

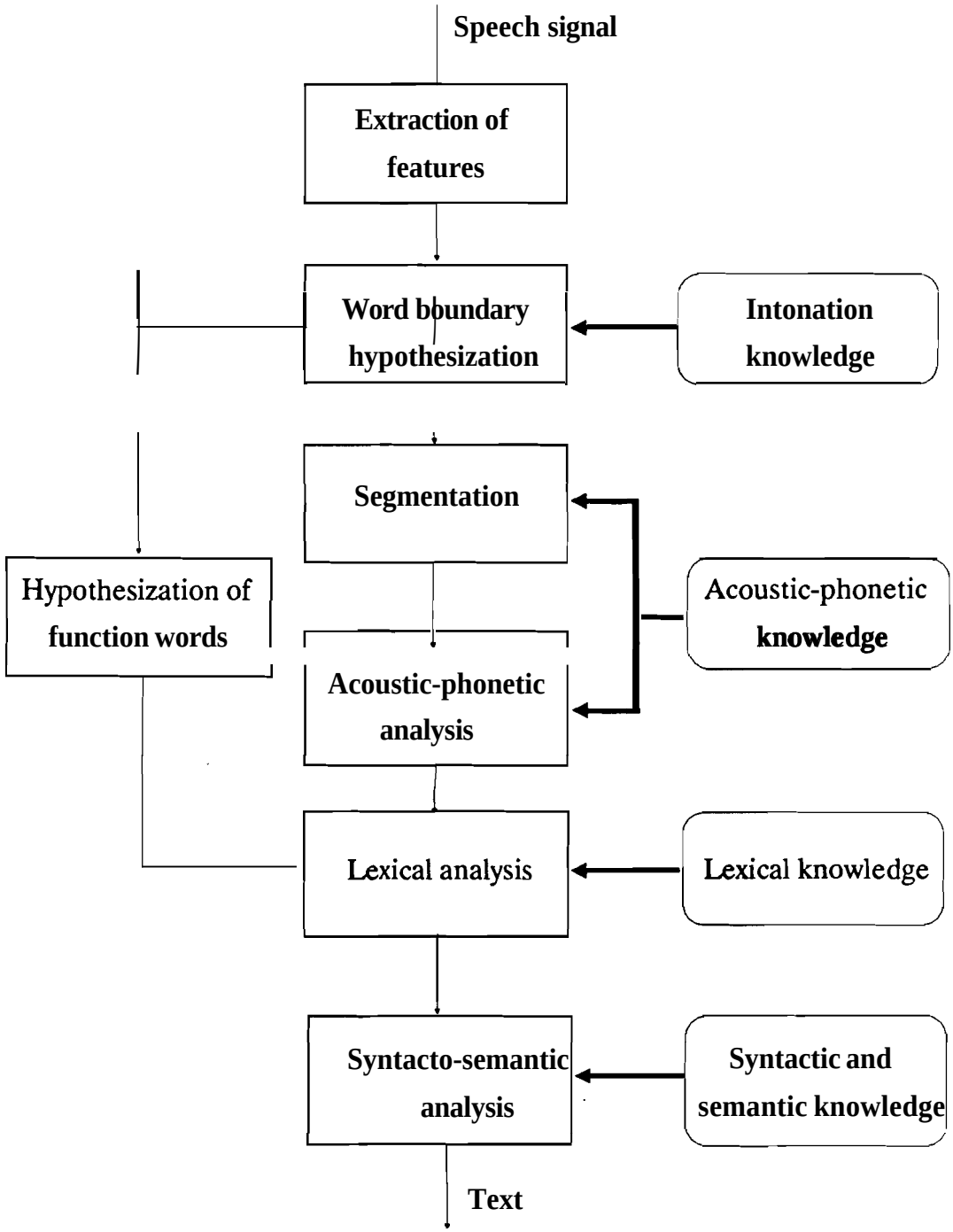


Fig.7.1. Block diagram of a speech-to-text system for Hindi

presupposes the availability of output symbols from the acoustic-phonetic module. Also, these modules have to wait for input until the acoustic-phonetic module completes its task of recognizing the symbols. Algorithms for hypothesizing word boundaries from phonological constraints and dictionary matching take much time. They may also give unreliable results since these algorithms have to process the phonemic/character lattice generated due to ambiguity in the output of the acoustic-phonetic decoding. It is also possible that several alternative word sequences may result due to several possible legal combinations of the basic units in the lattice. The hypothesized words have to be passed to syntactic and semantic analyzer modules to eliminate the wrong alternatives. Thus the errors produced by the acoustic-phonetic module due to variability in the input speech propagate through the higher modules. Word boundary hypothesization, in this context, based on some acoustic parameters which are correlates of prosodic features is not affected by these errors.

The chapter is organized as fallows: Section **7.2** discusses pitch accent features derived from the pitch accent patterns of continuous speech in Hindi. An algorithm to hypothesize word boundaries from continuous speech in Hindi based on the pitch accent features is discussed in Section **7.3.** It also discusses the results of this algorithm and analyzing the errors. Section **7.4** discusses an algorithm to hypothesize function words from continuous speech in Hindi using word boundary hypothesizationalgorithm.

7.2 Deriving pitch accent features from pitch accent patterns in Hindi

We have described the F_0 contour within a prosodic word in terms of valleys and peaks in Section **3.3.2.** Valleys and peaks correspond to the prominence of a particular syllable in the content word. Monosyllabic words are considered as a unique category in which both valley and peak occur in the same syllable. In disyllabic and trisyllabic words, the initial syllable corresponds to the valley and the final syllable corresponds to the peak. This assignment ignored the status of intermediary syllables in the trisyllabic, tetrasyllabic and pentasyllabic words. The status of the intermediate syllables are predictable usingpitch accent

features. Pitch accent feature is defined as the prominence of a syllable with respect to the immediately preceding syllable as reflected by the value of the pitch frequency.

In order to describe the pitch accent patterns in Hindi we use two features, namely, *Low* (L) and High (H) which are decided by the relationship of F_0 value of nucleus of the syllable (vowel) with respect to the F_0 value of nucleus of the immediately preceding syllable in an utterance such that $F_0(H) > F_0(L)$. The advantages of selecting nucleus of syllable are (1) pitch extraction algorithm gives reliable results at syllable nucleus; (2) pitch values do not change significantly even if the speech is collected in noisy input conditions because we consider only the high signal-to-noise ratio (SNR) portions of speech and (3) the mid point of syllable nucleus is less affected by the coarticulatory phenomena. The pitch accent feature for the syllables in a disyllabic word is LH. Consider the case of a trisyllabic word. All syllables in a word are necessarily assigned to a feature, L or H. Pitch accent for the initial syllable is L and for the final syllable is H. The Fo value of the second syllable is higher than the first syllable and hence the pitch accent is H. In order to avoid ambiguity, syllables of the same category are marked by ascending indices of L_i or H_i. Thus the pitch accent feature for a trisyllabic word is represented by LH_1H_2 where $F_0(H_2) > F_0(H_1)$.

Pitch accent of a monosyllabic content word can be considered as a unique case. The F_0 value of such a syllable nucleus may be less than the F_0 value of the previous syllable nucleus. Hence the pitch accent pattern of a monosyllabic word would be assigned L. But for distinguishing a monosyllabic content word from a function word, we use the valley-peak assignment of monosyllabic content words. As discussed earlier, F_0 pattern of a monosyllabic content word shows a valley followed by a peak within the same syllable. Monosyllabic word with this Fo pattern is hypothesized as content word and the pitch accent pattern is assigned as H_0.

Fig.7.2 shows some examples of pitch accent features in Hindi. Fig.7.2a is a disyllabic word with a pitch accent feature of LH. Fig.7.2b is a monomorphemic tetrasyllabic word. As per local fall-rise patterns, valley occurs at the initial

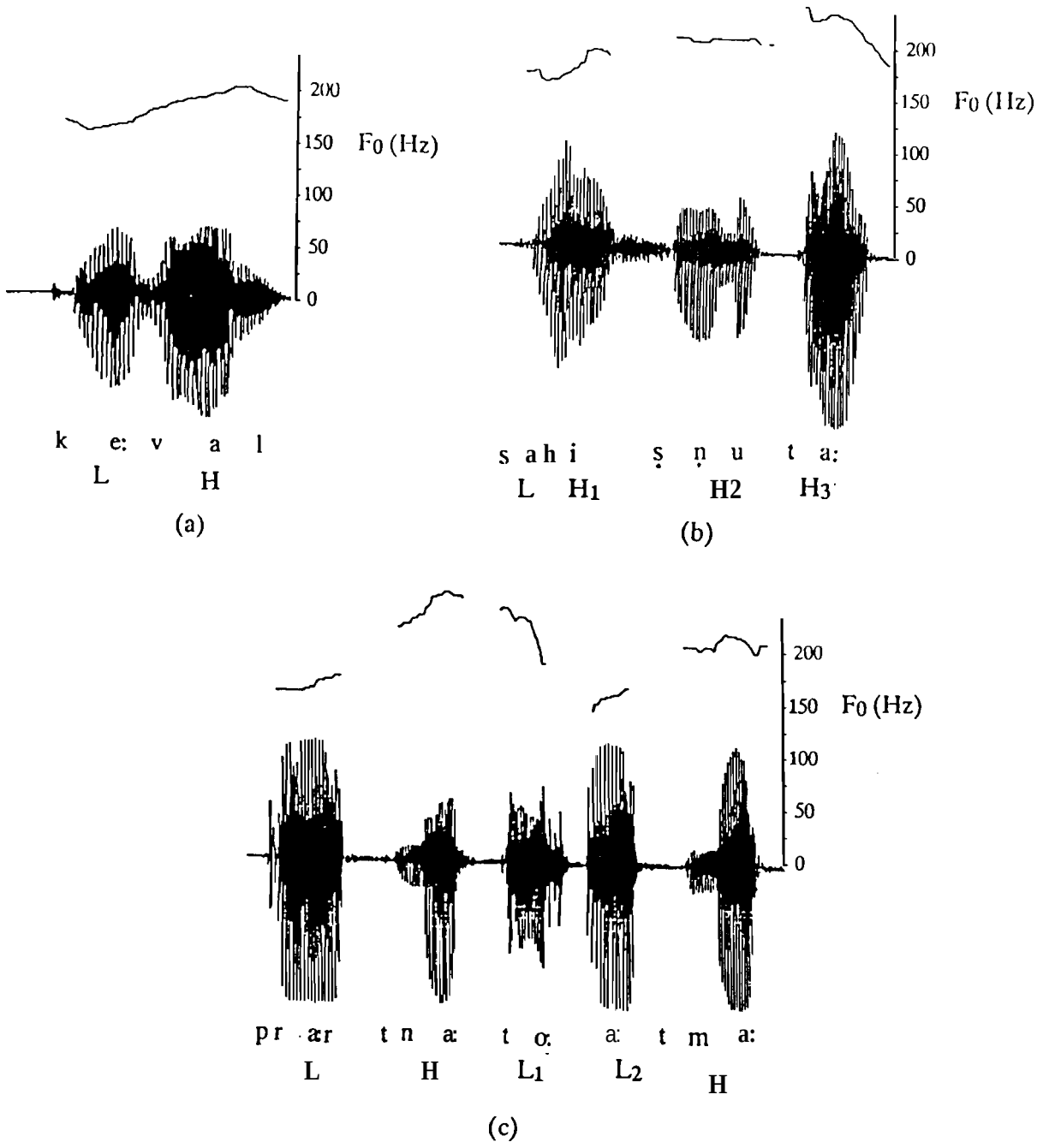


Fig.7.2. Examples of pitch accent features in Hindi

(a) Disyllabic word *[ke:val]* (absolute) (LH)

(b) Monomorphemic tetrasyllabic word *[sahiṣṇuta:]* (tolerance) ($LH_1H_2H_3$)

(c) Two disyllabic word with an accented function word in between *[pra:rtna:#to:#a:tma:]* (LHL_1L_2H). # indicates a word boundary.

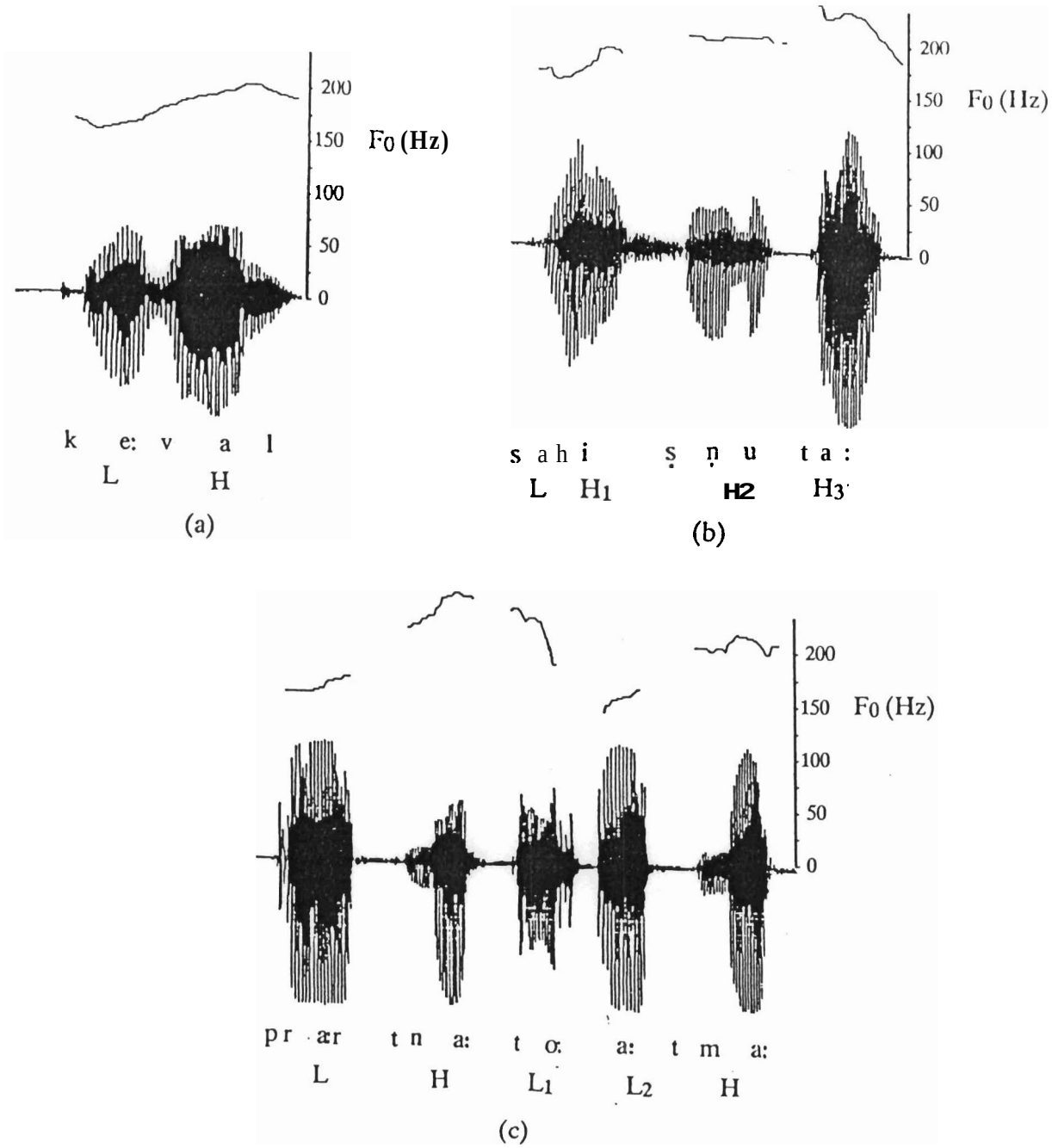


Fig.7.2. Examples of pitch accent features in Hindi

(a) Disyllabic word *[ke:val]* (absolute) (LH)

(b) **Monomorphemic tetrasyllabic word *[sahiṣṇuta:]* (tolerance)** ($LH_1H_2H_3$)

(c) **Two disyllabic word with an accented function word in between *[pra:rtna:#to:#a:tma:]*** (LHL_1L_2H). **# indicates a word boundary.**

syllable and peak at the final syllable. The corresponding pitch accent feature is $LH_1H_2H_3$. Fig.7.2c is an utterance with two disyllabic words with an unaccented function word in between. Corresponding pitch accent feature is LHL_1L_2H.

Each syllable in the utterance of a word has a pitch accent feature of either L_i or H_i depending the number of syllables in a word. From our data, we have found that the maximum number of syllables possible for a monomorphemic word is four. The assignment of features L_i or H_i for each syllable is done in such a way that

$$F_0(L_{i+1}) < F_0(L_i) \text{ and}$$
$$F_0(H_{i+1}) > F_0(H_i) \text{ where } 0 < i < 4.$$

The syllable with the highest index in a sequence of identical features (L or H) coincides either with a valley or with a peak. For a trisyllabic word the pitch accent pattern is LH_1H_2 and its peak is characterized by $H2$ and the valley by L.

Table 7.1 summarizes the terminology used in the word boundary hypothesization algorithm.

Terms	Parameters to describe the term
(1) Intonational phrase	F_0 pattern of sentence
(2) Targets of F_0	(i) Valley(V)
	(ii) Peak (P)
(3) Pitch accent pattern	(i) Low (L_i, $1 \leq i < 4$)
	(ii) High (Hi, $1 \leq i < 4$)

Table 7.1. The terminology used in the word boundary hypothesization algorithm

The relation between Fo targets and the corresponding pitch accent features for words with different syllabic compositions are given in Table 7.2.

The position and the type of a monosyllabic function word affect the pitch accent pattern significantly. The classification of such function words and their pitch accent behavior are discussed in Section 3.3.2.1. The pitch accent pattern

Type of word	Fo targets	Pitch accent pattern
Monosyllabic word	**V P**	H_0
Disyllabic word	**V P**	**L H**
Trisyllabic word	**V P**	**L H_1 H_2**
Tetrasyllabic word	(i) **V** P	**L Hi H_2 H_3**
	(ii) **V** P V P	**L H L H**
Pentasyllabic word	**VP VP**	(i) **L H L H_1 H_2**
		(ii) **L Hi H2 L H**

Table 7.2. The relation of F_0 targets and pitch accent patterns for content words

for a disyllabic content word followed by a prosodically cohesive (accented) function word may be similar to that of a trisyllabic word, and hence the pitch peak occurs on the following function word. The reassigned pitch accent feature for the prosodic word is therefore given by

 # L H # H # ----> # L H_1 H_2 #

where # represents word boundary.

The boundaries of disyllabic and trisyllabic prosodic words in continuous speech are hypothesized by identifying the **H** and **H2** respectively. That is, we hypothesize a word boundary after the largest indexed **Hi** in the sequence. For example, a monomorphemic tetrasyllabic word (or a trisyllabic word with following accented monosyllabic function word) has the pattern $LH_1H_2H_3$ and hence a boundary is hypothesized after **H3.** Instances of a tetrasyllabic word with the following accented function word were not found in our database. In case of such a possibility, there would be either (1) a single boundary following H_4 (peak), or (2) two boundaries including a morphemic boundary. For polymorphemic words we would get an error in the output of the word boundary hypothesization, as morphemic boundary is not an expected output. Such a division can be avoided if morphemic information is available. But acquisition and representation of morphemic knowledge for word boundary hypothesization

is a different problem. However, these errors do not cause serious problems in applications, partly because such occurrences are rare, and also partly because such a boundary coincides with morphemic boundary.

7.3 An algorithm for hypothesizing word boundaries using pitch accent features in Hindi

We have developed an algorithm based on the pitch accent features in Hindi to hypothesize word boundaries. The algorithm can be divided into three parts. The first part detects intonational phrases in the utterance using pauses and resetting of F_0 contour. Then the algorithm detects peaks in the energy contour, which correspond to the syllable nuclei. The final part of the algorithm hypothesizes word boundaries based on pitch accent patterns. In continuous speech a word boundary can be hypothesized between syllables whose pitch accent feature is changing from H_i with highest value of i to L. That is, a word boundary can be hypothesized between two syllables if F_0 value of the first syllable is greater than F_0 value of the second syllable. Thus first step of the algorithm compares pitch accent features of the syllables and hypothesizes word boundaries. The final part of the algorithm takes care of a unique case in which the pitch accent feature Hi rises beyond four syllables. The assumption behind this is the maximum number of syllables possible for a monomorphemic word is four. The corresponding pitch accent pattern is $LH_1H_2H_3$. Hence we can hypothesize a boundary after the syllable which has a pitch accent feature Hi where i is greater than three. The algorithm for word boundary hypothesization is given in Fig.7.3. Fig.7.4 shows an example of hypothesization of word boundaries using this algorithm.

The advantage of the proposed algorithm for word boundary hypothesization is that it works even under noisy speech input conditions. The robustness of the algorithm is primarily because only gross parameters like energy and pitch are used in its implementation. Also, in order to hypothesize word boundaries we considered pitch accent features which are derived from high SNR portions of speech. The algorithm works fast since it uses the F_0

Let m be the number of intonational phrases and n be the number of syllables in an utterance.

Parts:

I. Identify intonational phrases by detecting silence region in the utterance
   using a threshold for the energy contour
   If duration (silence) > 300 msec then identify it as a pause
   Let Pi, P2,...P_{m-1} be the pauses
   Split the utterance into intonational phrases $C_1, C_2...C_m$ using the pauses

II. Apply 7 point median smoothing to energy contour
   Find peaks in the energy contour
   Avoid spurious peaks by checking the corresponding pitch values

III. For each intonatinal phrase Cj (for $j = 1, 2, ...$ m) do
   begin
      Using peaks in energy contour spot syllable nuclei
      Let Mi, M2...M_n are the midpoints of the syllable nuclei
      for i : = 1 to n-1 do
      begin
         Steps:
         1. if $F_0(M_i) > F_0(M_i+1)$ then place a word boundary in the midway between Mi and Mi+1
         2. if (i > 3) then
            if $(F_0(M_{i-3}) < F_0(M_{i-2}))$ and $(F_0(M_{i-2}) < F_0(M_{i-1}))$ and $(F_0(M_{i-1}) < F_0(M_i))$ then place a word boundary in the midway between Mi and Mi+1
      end
   end.

Fig. **7.3.** Algorithm for hypothesizing word boundaries from continuous speech in Hindi.

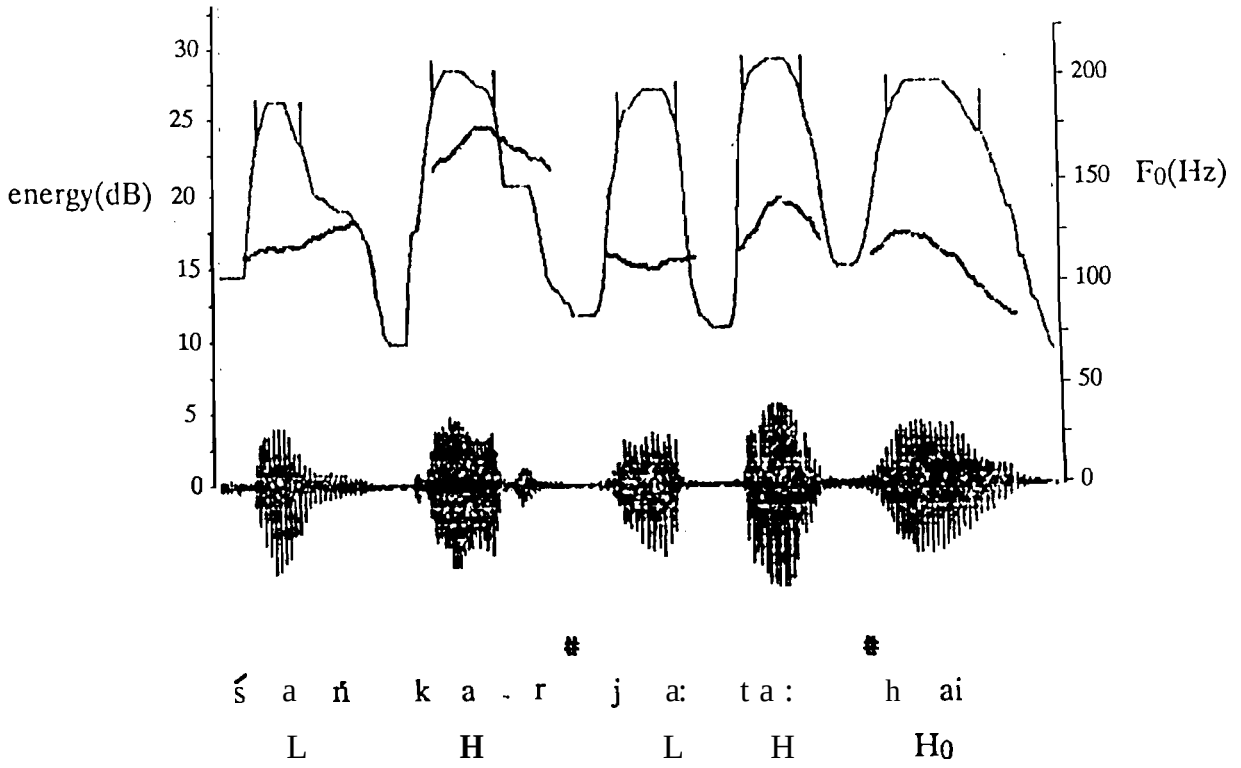


Fig.7.4. Hypothesization of word boundaries for the natural utterance corresponding to the sentence */śaṅkar#ja:ta:#hai/* (Shankar goes). Thick line indicates the pitch contour for the utterance and thin line corresponds to the energy contour. Vertical bars in energy contour indicate the location of syllable nuclei. # indicates a word boundary. L and H indicate pitch accent feature of each syllable.

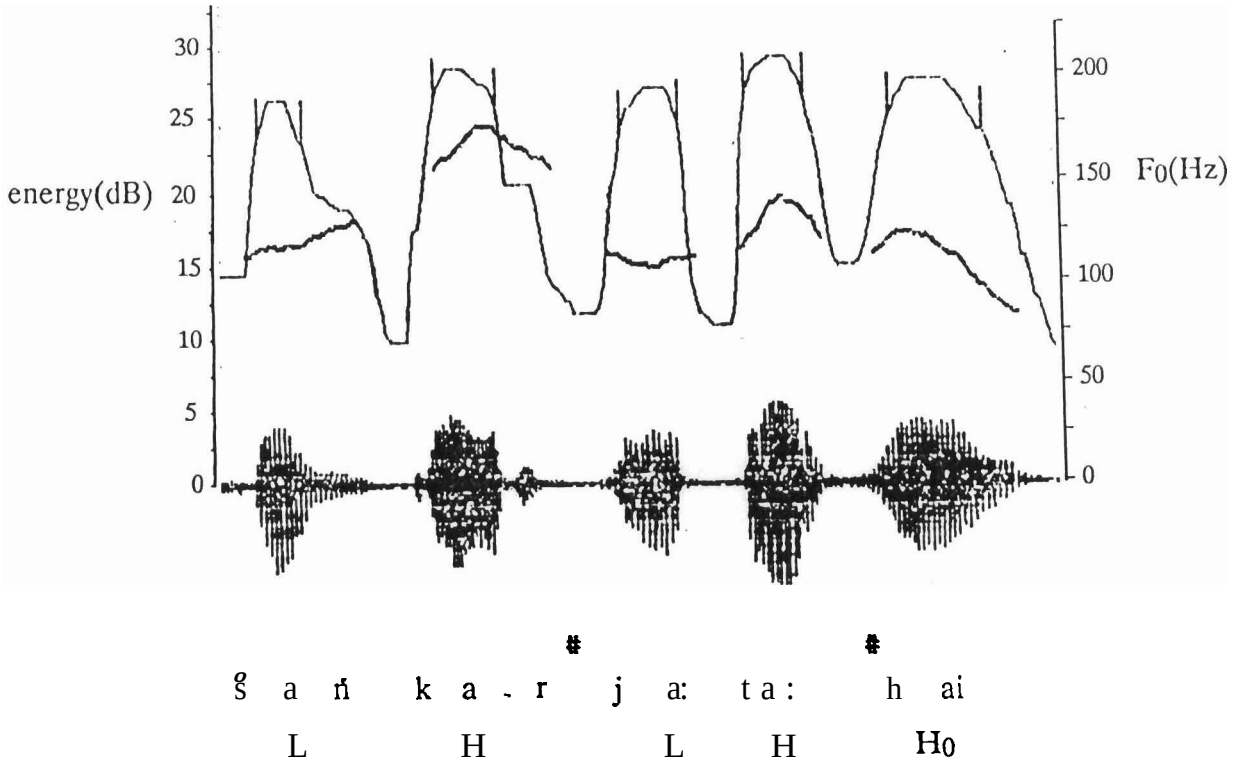


Fig.7.4. Hypothesization of word boundaries for the natural utterance corresponding to the sentence */ʃaɲkar#ja:ta:#hai/* (Shankar goes). Thick line indicates the pitch contour for the utterance and thin line corresponds to the energy contour. Vertical bars in energy contour indicate the location of syllable nuclei. # indicates a word boundary. L and H indicate pitch accent feature of each syllable.

values of syllable nuclei only.

73.1 Results and discussion

The word boundary hypothesization algorithm was evaluated on a corpus of data consisting of (a) a Hindi text of 2281 words (140 sentences) read out by two speakers and (b) a text of 264 words (30 sentences) from news bulletin in Hindi broadcast over All India Radio. The corpus of test data is a subset of data collected for the analysis of the properties of intonation patterns in Hindi speech. The sentences are of different lengths and they have different syntactic complexities. Speech was recorded in an ordinary office environment.

7.3.1.1 Performance in clean speech

Table 7.3 shows the performance of the word boundary hypothesization algorithm. HBHA (a speaker with Uttar Pradesh accent) and HCHO (a speaker with Bihar accent) are two native speakers of Hindi with different dialects. HNEWS is a part of Hindi news bulletin broadcast over All India Radio. The dialect used for broadcast over All India Radio is generally considered as the standard variety of Hindi. The speaking rate was measured by counting the number of words spoken per minute. Accuracy rate is calculated as the ratio of number of correct boundaries hypothesized to the total number of word

Speaker	Size of corpus (in words)	Speaking rate (wpm)	Accuracy rate %	Error rate %
HBHA	1471	157.72	83.12	13.12
HCHO	810	132.60	78.77	16.79
HNEWS	264	160.01	70.08	11.74
Avg		450.11	77.32	13.88
Total	2545			

Table 7.3. Results of the word boundary hypothesization algorithm

boundaries. The error rate is the ratio of total number of incorrectly hypothesized boundaries to the total number of hypothesized boundaries. The algorithm hypothesized an average of 77.32% of boundaries correctly. This accuracy was maintained more or less the same for all speakers. 13.88% of hypothesized boundaries were wrong. The performance figures show that the pitch accent pattern is free from the dialect variations of Hindi and therefore speaker independent. The absolute F_0 value may vary from speaker to speaker depending on various factors such as age, sex, etc.. But the word boundary hypothesization algorithm depends upon the pitch accent pattern and hence is independent of the absolute F_0 values. The algorithm is also vocabulary independent since the pitch accent pattern does not depend on any vocabulary.

7.3.1.2 Performance in noisy conditions

Since the ultimate objective is to hypothesize word boundaries in continuous speech for a speech-to-text system, the performance of the algorithm was evaluated for noisy speech as well. A list of 50 sentences were selected at random from the database and noise was added to make the overall signal-to-noise ratio 3dB. Fig.75 shows the performance of the algorithm on noisy speech. The pitch and energy contour of the noisy speech are also plotted. Voiced/unvoiced decision is not included in the pitch detection algorithm. The energy was computed as sum of squares of the samples in the frame. The parameters are slightly disturbed due to the addition of noise. But the performance of the algorithm is not altered by the noise, because the pitch accent pattern remained more or less the same. It is obvious from Fig.7.5 that the algorithm hypothesized word boundaries well even for the noisy input conditions.

Table 7.4 shows comparison of the performance of the algorithm for noisy and clean speech. It is interesting to note that the accuracy rate in noisy conditions is slightly higher (81%) than that of clean speech (77.32%). But the percentage of wrong boundaries is also increased (17% from 13.88%). This is due to errors in the detection of energy peaks from noisy speech signal.

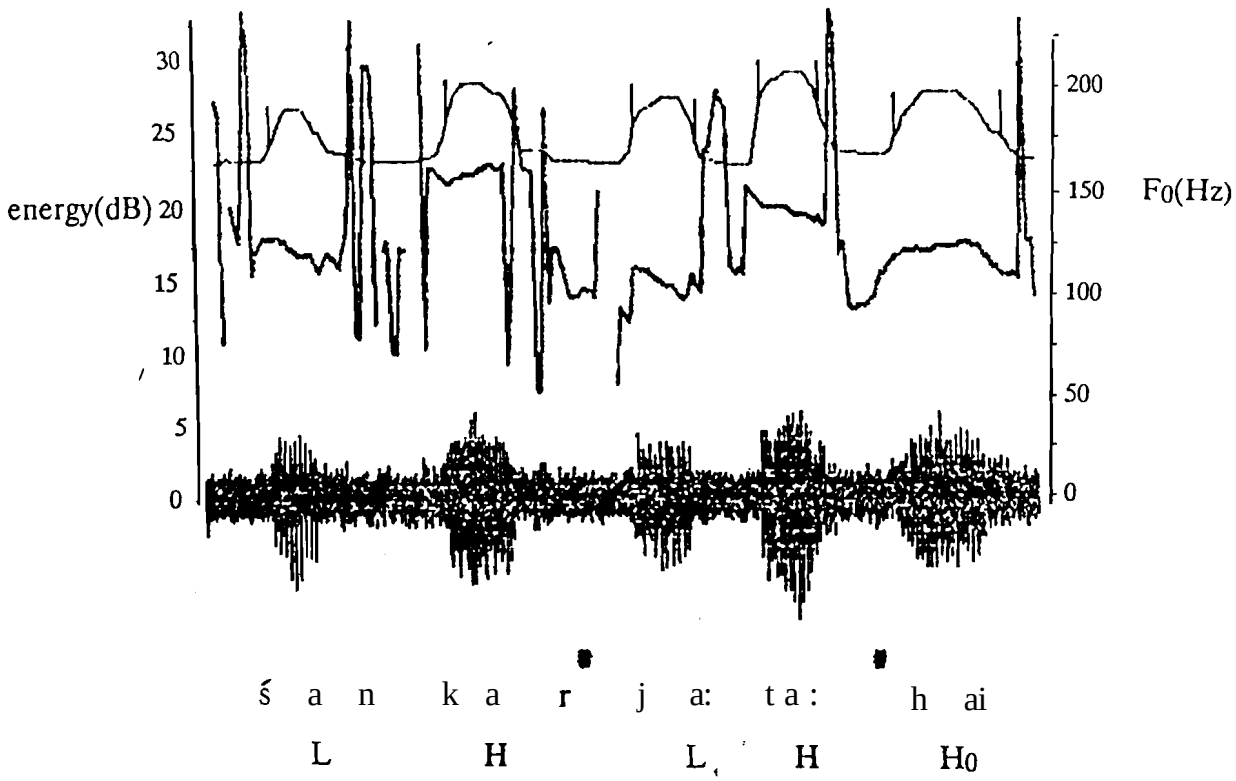


Fig.7.5. Hypothesization of word boundaries for adverse speech input conditions. The clean speech corresponding to the sentence */śaṅkar#ja:ta:#hai/* (Shankar goes) is mixed with random noise (SNR = 3dB). Thick line indicates the pitch contour for the utterance and thin line corresponds to the energy contour. No voiced/unvoiced decision is taken at the pitch estimation. Vertical bars in energy contour indicate the location of syllable nuclei. # indicates a word boundary. L and H indicate pitch accent feature of each syllable.

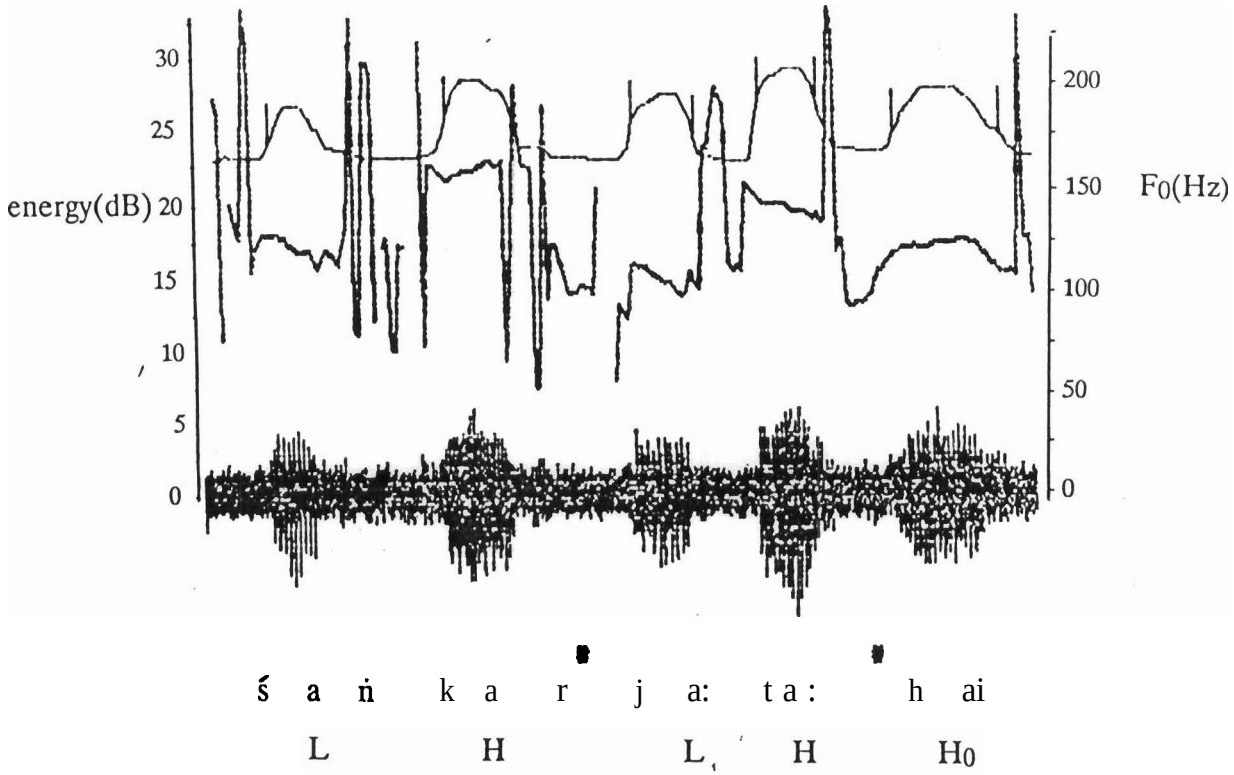


Fig.7.5. Hypothesization of word boundaries for adverse speech input conditions. The clean speech corresponding to the sentence */śaṅkar#ja:ta:#hai/* (Shankar goes) is mixed with random noise (SNR = 3dB). Thick line indicates the pitch contour for the utterance and thin line corresponds to the energy contour. No voiced/unvoiced decision is taken at the pitch estimation. Vertical bars in energy contour indicate the location of syllable nuclei. # indicates a word boundary. L and H indicate pitch accent feature of each syllable.

Corpus	% of correct boundaries	% Error rate
Clean Speech	77.32	13.88
Noisy Speech	81	17

Table 7.4. Robustness of the boundary detection algorithm

73.2 Analysis of errors

The occurrence of errors in the word boundary hypothesization can be classified into two categories: (1) incorrect boundaries placed where there are no boundaries and (2) undetected (missed) boundaries.

7.3.2.1 Errors due to incorrect boundaries

In our experiments on clean speech 13.88% of the hypothesized boundaries are wrong. Similarly, the percentage of wrong boundaries in noisy speech is 17%. Table 7.5 shows the break up of various categories of the wrong boundaries. The algorithm hypothesizes wrong boundaries due to several reasons such as absence of energy peaks, wrong values of F_0, low range of F_0 contour, wrong valley-peak prediction, coarticulatory influences and morphemic characteristics of tetrasyllabic and pentasyllabic words. Among these, the major cause for wrong

Speaker	Causes								Total
	a	b	c	d	e	f	g	h	
HBHA	0.75	1.19	3.06	0.25	3.06	2.06	0.91	1.84	13.12
HCHO	1.83	5.17	6.31	0.11	0.46	0.92	0.69	1.30	16.79
HNEWS	1.03	3.79	4.83	0.00	1.03	0.35	0.00	0.71	11.74
Avg	1.20	3.38	4.73	0.12	1.52	1.11	0.53	1.28	13.88

(Causes: (a) absence or wrong detection of energy peak (b) wrong F_0 values (c) low F_0 range (d) wrong valley-peak prediction (e) high peak on the previous syllable (f) low F_0 value on the following valley (g) tetrasyllabic words and (h) error caused by other factors)

Table 7.5. Analysis of Error boundaries

boundaries is the low range of F_0 contour (4.73%). The F_0 tapers off and the range of F_0 contour gets reduced towards the end of the clause boundary in general, and the sentence boundary in particular. The tapering effect towards the end of utterances sets off from the beginning of the word final syllable which has a pitch accent feature H. The range of F_0 contour between two points is small in some cases due to this tapering effect. Another cause of incorrect word boundaries is the wrong values of F_0 due to spurious peaks in the Fo contour (3.38%). Coarticulation affects the word boundary hypothesization in two ways - as a high peak of F_0 on the previous syllable or as a low valley of F_0 on the following syllable. 1.52% of the errors were caused by the high peak of Fo on the previous syllable. Coarticulatory influences on the low valley of F_0 resulted in an error of 1.11%. Emphasis of words are characterized by wider range of F_0 between the valley and peak. When the peak value of F_0 of a content word is very high, the actual valley of the following syllable would be on the right part of the syllable nucleus. Hence, the error is caused when the algorithm takes into consideration only the midpoint values. In continuous speech the vowels in sequence (eg./*hua:*/, /hue:/) get merged. In such instances syllable segmentation based on energy may miss one of the syllables, Such errors constitute 1.20%. Bimorphemic tetrasyllabic words with pitch accent pattern of LHLH resulted in an error of 0.53%. It is interesting to note that such boundaries correspond to morphemic boundary. 0.12% of the errors are caused by wrong prediction of valleys or peaks. 1.28% of the errors are caused by other reasons such as wrong detection of syllable nucleus, etc..

7.3.2.2 Errors due to undetected boundaries

Table 7.6 shows the analysis of errors for the undetected boundaries. In the overall performance of the algorithm, the ratio of the undetected boundaries is higher than the ratio of the incorrect boundaries. In clean speech 22.68% of the boundaries were missed. For noisy speech 19% of the boundaries were undetected. The major portion of the undetected boundaries (14.88%) are due to the function words which prosodically conjoin with the preceding or the

Speaker	Causes				Total
	a	b	c	d	
HBHA	**14.56**	**0.21**	**0.70**	**1.41**	16.88
HCHO	**12.44**	**4.03**	**4.37**	**0.39**	21.23
HNEWS	**17.64**	**0.39**	**9.59**	**2.30**	29.92
Avg	**14.88**	**1.54**	**4.89**	**1.37**	**22.68**

(Causes: (a) function word conjoins with the following or preceding content word, (b) absence of energy peaks, (c) absence of Fo or spurious Fo at the mid point of the vowel, and (d) other reasons)

Table **7.6.** Analysis of **Undetected(missed) WBs**

following noun phrase. The errors in pitch estimation accounted for **4.89%** errors. It is due to the absence of F_0 or spurious F_0 at the midpoint of the vowel. The algorithm detects the midpoint of the syllable nucleus based on peaks in the energy contour. The absence of energy peaks resulted in **1.54%** errors. Various other causes account for **1.37%** missed boundaries.

7.4 Hypothesizing function words in continuous speech in Hindi using word boundary hypothesization algorithm

In this section we show how function words of a language can be hypothesized using pitch accent patterns. Function words are a class of words which bear only grammatical information. These are different from the content words which are semantically important. We can hypothesize monosyllabic function words by using pitch accent features. For this we need to have an idea of the occurrence and nature of function words in a language. The pitch accent patterns of function words in Hindi were discussed in Section **3.3.2.**

There are three types of pitch accent patterns for monosyllabic function words. Function words like relative pronouns and conjunctions have independent existence. Their pitch accent pattern is same as that of monosyllabic content words (Ho). These function words cannot be hypothesized using pitch accent features alone. All other function words may conjoin with the preceding

or the following noun phrase. Among these, some function words (e.g., post position */ko:/*) may get accented and therefore prosodically they conjoin with the previous content word. In such cases pitch accent pattern at the function word will have the highest value of Hi. As discussed in Section 7.2, the pitch accent pattern for a disyllabic word (LH) with the following accented function word (H) will be LH_1H_2. A third category of pitch accent pattern occurs when the function word is in an unaccented position. Such function words can be hypothesized using word boundary hypothesization algorithm.

Consider the case of an unaccented monosyllabic function word placed in between two disyllabic words. Disyllabic words have the pitch accent pattern LH. Since the function word is unaccented, the F_0 at the function word will be less than the peak F_0 value of the previous disyllabic word. Since the pitch accent is defined as the relative prominence of a syllable with respect to the preceding syllable, the pitch accent feature of the unaccented function word is assigned L As discussed in Section 2, the pitch accent pattern is reassigned as follows:

$$L\ H\ L\ L\ H \text{ ----> } L\ H\ L_1 L_2\ H.$$

where, L and L_2 correspond to valleys and H corresponds to the peak. The pitch accent feature L_1 corresponds to an unaccented function word. It is possible to hypothesize that in the sequence $L_1\ L_2$, the L_1 must be for a function word. Hence, the above pitch accent pattern suggests the following hypothesization of boundaries:

$$L\ H\ L_1\ L_2\ H \text{ ----> } \#\ L\ H\ \#\ L_1\ \#\ L_2\ H\ \#$$

where # represents word boundary and the feature L_1 corresponds to a function word.

In our study, we have found that prosodically cohesive function words belonging to the left-occurring category (function words preceded by a noun phrase) are very difficult to hypothesize. In such cases the final peak of the prosodic word usually falls on the nucleus of the following monosyllabic function word and its pitch accent feature is the highest Hi. The possible location for a monosyllabic function word in a pitch accent pattern is either L_i with the lowest value of i or Hi with the highest value of i. In the case of more than one function

word in a sequence, the last function word is prosodically conjoined with the following word. In a speech-to-text system this method can either hypothesize the function word in an utterance or restrict the search space by finding the possible syllabic slots (either L_i with the lowest i or H_i with the highest i) in the pitch accent pattern for function words. This method can also enhance the capability of the acoustic-phonetic module and can reduce the complexity in the lexical access in a speech-to-text system for Hindi.

7.4.1 An algorithm for hypothesizing function words in Hindi

The pitch contour for the syllable corresponds to the unaccented function word decreases monotonically from the peak of previous content word to the valley of the next content word. We use this knowledge together with word boundary hypothesization algorithm for hypothesizing some function words in Hindi.

Algorithm for hypothesizing function words in Hindi based on word boundary hypothesization algorithm is given in Fig.7.6. The algorithm contains two parts. Initially the word boundaries are hypothesized using word boundary hypothesization algorithm. The monosyllabic words hypothesized using the algorithm are analyzed separately to find out the position of unaccented monosyllabic function words. That is, as discussed earlier, in a pitch accent pattern of L_1L_2, L_1 corresponds to a monosyllabic word. F_0 contour for an unaccented function word is monotonously decreasing and hence the slope of F_0 contour at syllable nucleus is negative whereas for a monosyllabic content word Fo contour exhibits a valley and a peak at the same syllable. In order to separate function words from monosyllabic content words, we consider the slope of F_0 contour at the syllable nucleus. If slope is negative, then hypothesize the corresponding monosyllabic word as function word.

7.4.2 Results and discussion

Fig.7.7 shows some speech waveforms and the corresponding actual and hypothesized word boundaries. From Figs.7.7a-d it is seen that some of the function words are hypothesized (marked as fw) by the algorithm. All these

Let m be the number of intonational phrases and n be the number of syllables in an utterance.

For each intonatinal phrase Cj (for $j = 1, 2, \ldots m$) do

begin

Hypothesize word boundaries using word boundary hypothesization algorithm

Let Mi, $M_2 \ldots M_n$ are the midpoints of the syllable nuclei

for i := 1 to n-1 do

begin

If ($i > 1$) then

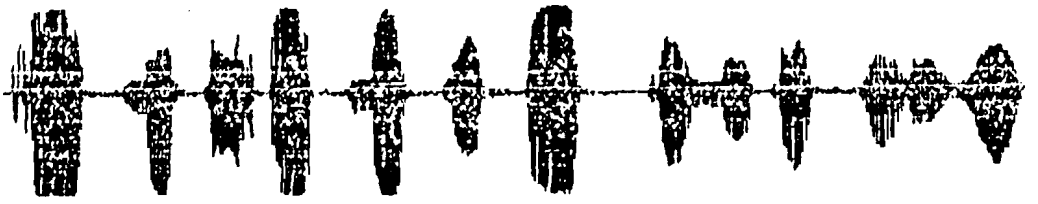
if $(F_0(M_{i-1}) > F_0(M_i))$ and $(F_0(M_i) > F_0(M_{i+1}))$ then mark the syllable corresponding to Mi as a monosyllabic word.

If ((Mi corresponds to monosyllabic word) and (slope of F_0 at Mi < 0)) then the syllable corresponding to Mi is a monosyllabic unaccented function word

end

end.

Fig.7.6. Algorithm for hypothesizing function words in Hindi from continuous speech using word boundary hypothesization algorithm.



4

fv fv fv

pr a:r t n a: to: a: t m a: k o: s a:f k a r n e: ka: j a: ḍ u: h ai

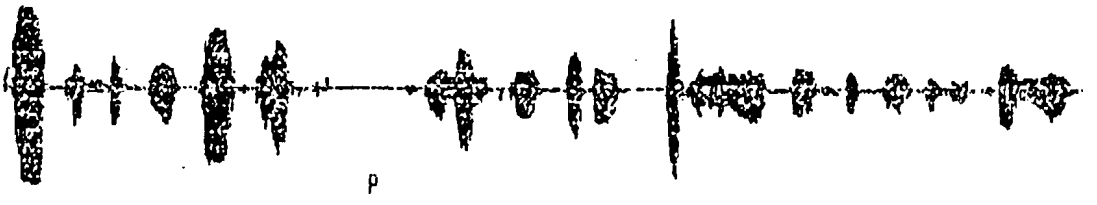
(a)

p

fv fv

j o: ki s i: se: i: r ṣ y a na hĩ:ra ht a: j *o:* da ya: k a: ban da:r h ai

(b)



p

fv

a:t maśudhi: k a: ar th j i: van ki: sabhi: pahalõ: se: śu dhi: ho:na:ca:hiye:

(c)

fv fv fv

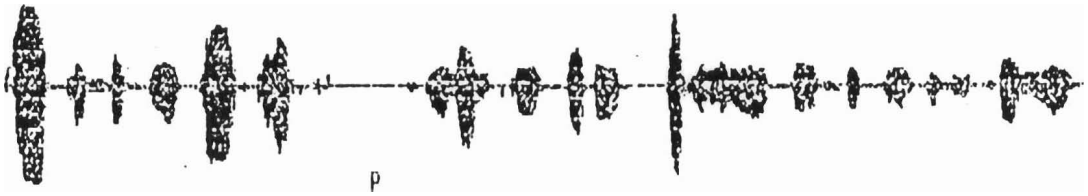
pr a:r t n a: to: a: t m a : k o : s.a: f k a r n e : ka: j a: ḍ u: h ai

(a)

fv fv

j o: ki s i: se: i: r ṣ y a na hĩ:ra ht a: j o: da ya: k a: ban da:r h ai

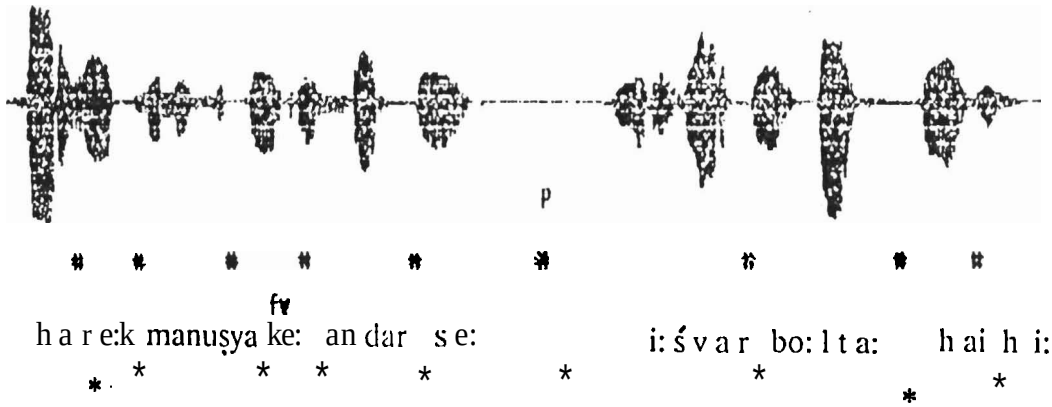
(b)



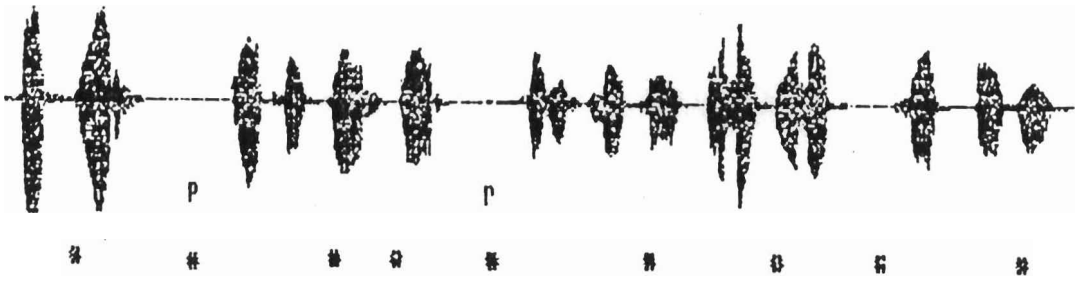
fv

a:t maśudhi: k a: ar th j i: van ki: sabhi: pahalõ: se: śu dhi: ho:na:ca:hiye:

(c)



(d)



(e)

Fig.7.7. Hypothesization of word boundaries for continuous speech corresponding to the following Hindi sentences.

(a) */pra:rtna: ko: a:tma:* ko: *sa:f kame: ka: ja:ḍu: hail*

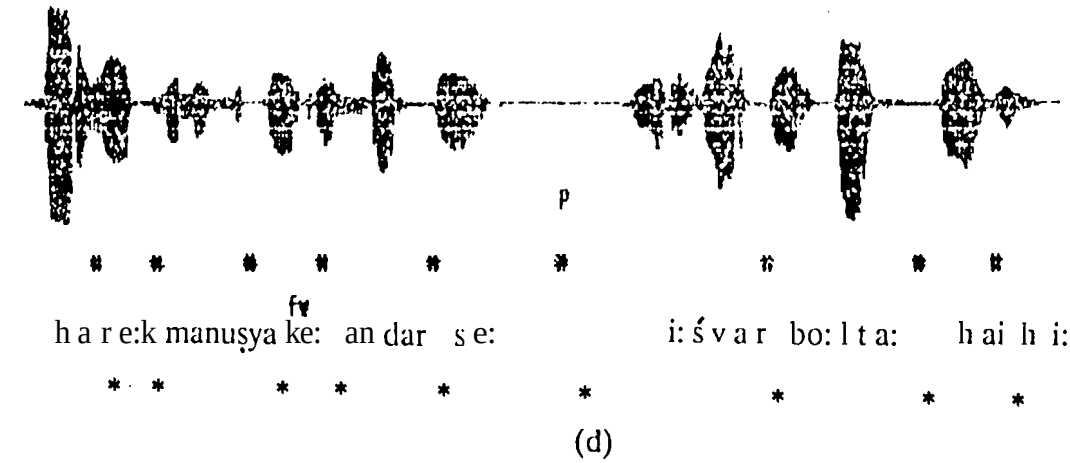
(b) */jo: kisi: se: i:ṛṣya nahĩ: rahta: jo: daya: ka: bhanḍa:r hai/*

(c) */a:tma śudhi: ka: arth ji:van ki: pahlõ: se: śudhi: ho:na: ca:hiye:/*

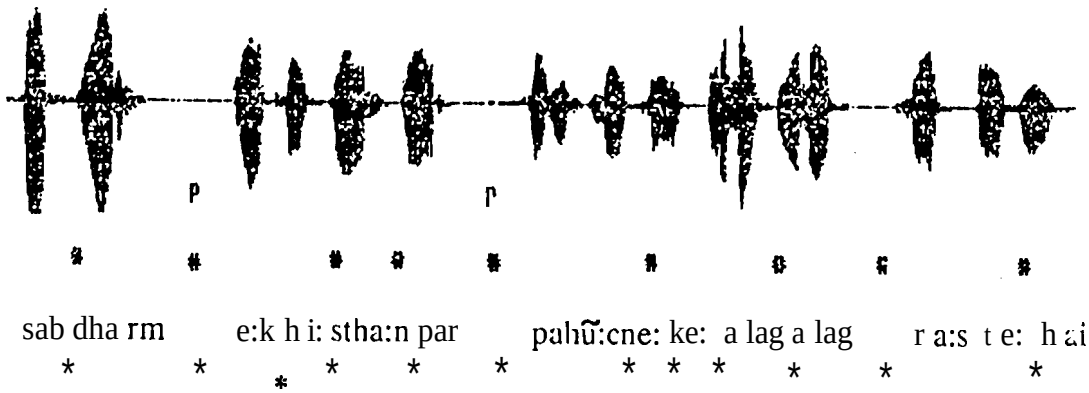
(d) *lhar e:k manuṣya ke: andar se: i:śvar bo:lta: hai hi:/*

(e) */sab dharm e:k hi: stha:n par pahũ:c ne: ke: alag alag ra:ste: hail*

The algorithm detected word boundaries and located function words. Here, P indicates pause corresponding to syntactic boundaries, # indicates the word boundaries hypothesized by the algorithm, * indicates actual word boundaries and fw corresponds to the hypothesized function words.



(d)



(e)

Fig.7.7. Hypothesization of word boundaries for continuous speech corresponding to the following Hindi sentences.

(a) */pra:rtna: ko: a:tma:* **ko:** *sa:f* **kame:** *ka: ja:ḍu:* **hail**

(b) */jo: kisi:* **se:** *i:ṛṣya nahĩ: rahta:* **jo: duya: ka:** *bhanḍa:r* **hail**

(c) */a:tma śudhi:* **ka:** *arth ji:van* **ki:** *pahalõ:* **se:** *śudhi: ho:na: ca:hiye:/*

(d) ***lhar e:k*** *manuṣya* **ke:** *andar* **se:** *i:śvar bo:lta:* **hai** *hi:/*

(e) */sab dharm* **e:k hi:** *stha:n par pahũ:c* **ne: ke:** *alag alag ra:ste:* **hail**

The algorithm detected word boundaries and located function words. Here, P indicates pause corresponding to syntactic boundaries, # indicates the word boundaries hypothesized by the algorithm, * indicates actual word boundaries and fw corresponds to the hypothesized function words.

function words are in unaccented positions. Accented function words are not hypothesized by this algorithm. As discussed earlier, function words in accented position conjoin with the previous content word and is difficult to separate. Consider Fig.7.7e. It contains monosyllabic function words */hi/* (*/e:k* hi/), */-ne:/* (*/pahũ:c ne:/*) and */ke:/* (*/ke: alag/*), but failed to hypothesize any of their occurrence because these function words prosodically conjoined with the preceding or the following content words. In such cases we have to find out other methods to isolate the function words. But in all these cases the syllable location of a function word is obvious - either the initial or the final syllable of a prosodic word.

7.5 Summary

In this chapter we proposed an algorithm to hypothesize word boundaries from continuous speech in Hindi based on pitch accent features. Pitch accent features were derived from the pitch accent patterns in Hindi. The word boundary hypothesization algorithm was tested on a large speech data. The algorithm performed well even in adverse speech input conditions such as noisy speech. The word boundary hypothesization algorithm can be used as an effective front end in a continuous speech recognition system to supplement the acoustic-phonetic analysis. The algorithm can help to reduce complexity in lexical access to a large extent. Errors in the word boundary hypothesization were analyzed and categorized. Performance of the word boundary hypothesization algorithm can be improved further by incorporating other prosodic knowledge sources such as duration and intensity, and the knowledge related to coarticulation and morphology. The properties of pitch accent features were exploited to hypothesize unaccented monosyllabic function words in continuous speech.

Chapter 8
DESIGN AND DEVELOPMENT OF A SPEAKER RECOGNITION SYSTEM BASED ON INTONATION KNOWLEDGE

8.1 Introduction

The ability of human listeners to identify speakers from their voices is well known. They identify reliably familiar voices from continuous speech. The performance decreases for unfamiliar voices and for short segments of speech when the duration is less than a second. It has been shown that sometimes the accuracy of automatic speaker recognition exceeds that of humans for short test utterances (Pollack, Pickett & Samby, 1954; Reich & Duke, 1979). Hence speaker recognition is one area of artificial intelligence where machine may surpass human performance. This is especially true for unfamiliar speakers, where humans take long time to learn the new voice compared to that for machines. Also, the number of unfamiliar voices that can be retained in the short term memory of human brain is limited. This chapter discusses a method for identifying speakers using neural networks from the utterances of a fixed text based on the intonation knowledge.

Automatic speaker recognition systems are classified into speaker verification and speaker identification systems based on the method of recognition (O'Shaughnessy, 1986; Rosenberg, 1976). Automatic speaker verification compares the test pattern against the reference pattern of the claimed speaker and involves a binary (yes/no) decision of whether the test speech pattern matches the template of the claimed speaker. In automatic

speaker identification, the system chooses the best match for a test voice from the voices known to the system. Since the number of comparisons and decisions are equal to the number of voices known to the system, the error rate for a speaker identification system is likely to be greater than the error rate of a speaker verification system.

Speaker recognition can also be classified into closed set and open set identification based on the domain of application (Doddington, 1985). A closed set speaker identification refers to a domain where characteristics of speakers are prespecified and hence the number of speakers is limited. In contrast, an open set speaker identification has a possibility to add unknown voices at any time and hence the number of speakers is unlimited. The recognition accuracy for an open set speaker identification system is less compared to a closed set identification system due to lack of training. In the following discussion, the term speaker recognition refers to a text dependent closed set speaker identification, unless otherwise mentioned.

The block diagram of the speaker recognition system is given in Fig.8.1. It has two parts: one corresponds to feature extraction and the other corresponds to classification. Feature extraction includes processing of speech signal to get the parameters for speaker recognition. It consists of acquiring input speech signal, processing it for extraction of pitch contour, word boundary hypothesization and extraction of parameters. Classifier consists of a neural network based architecture which identify the speaker from these input parameters.

This chapter is organized as follows. Section 8.2 discusses the issues in extracting features for the speaker recognition. Section 8.3 discusses the architecture for the proposed speaker recognition systems. The results obtained from the systems are discussed in Section 8.4. This discussion covers the performance of the systems in noisy input conditions also.

8.2 Features for speaker recognition

One of the most important steps towards achieving a successful speaker recognition system is the selection of features of speech which are capable of

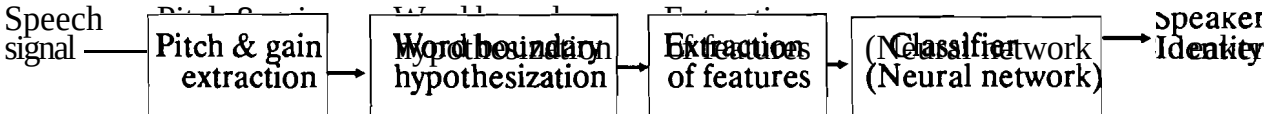


Fig.8.1. Block diagram of the proposed speaker recognition system. It consists of two parts, one corresponds to the feature extraction and the other for the classification. The features are pitch frequency values at valleys and peaks of pitch contour and durations of the words in the utterance of carrier sentence. Words are hypothesized using word boundary hypothesization algorithm with the help of pitch and energy contour. The classifier is a neural **netwok** based system in which network model is selected based on the domain of application.

representing speaker dependent properties in speech. Desirable characteristics of features of a suitable speaker recognition system can be summarized as below (Wolf, 1972; Atal, 1976): (1) they should be efficient in representing the speaker dependent information; (2) they should occur naturally and frequently in speech; (3) they should differ from one speaker to another; (4) extraction from speech should be easy; (5) they should not be susceptible to mimicry and (6) they should not be affected by reasonable background noise or transmission characteristics.

Two types of useful parameters can be derived from two different sources of speaker variability. They are: (1) the features related to the anatomical differences in the vocal tract (inherent features) and (2) the features related to the differences in the speaking habits (learned features). The anatomical differences relate to the fixed structural differences in the shape and size of the vocal tract which can vary considerably from one speaker to another. The differences in speaking habits are due to the manner in which speakers have learned to use their speech production mechanism. These differences indicate the temporal variations of speech characteristics of different individuals. A good example of such variations is the intonation pattern of continuous speech from different individuals.

Among the learned features, pitch contour is considered as a powerful feature for text dependent speaker recognition. Pitch contour has several advantages over spectral features in speaker recognition. Spectral patterns are affected by frequency characteristics of the recording and transmission systems. Also, spectral data depends upon the level at which speaker talks and the distance between speaker and the microphone. Pitch contour is unaffected by all these variations. Hence, it is one of the robust features for speaker recognition over other speaker dependent parameters.

In the proposed system, we consider the properties of pitch accent patterns for speaker recognition. The features used are the pitch frequency values of the valleys and peaks of the pitch contour of each prosodic word in the utterance together with duration of the words. From experiments, it is found that a person may be able to mimic the voice characteristics of a speaker which remains fixed

in time such as average pitch or time averaged spectrum. Mimicking entire pitch contour as a function of time appears to be difficult (Lummis & Rosenberg, 1972; Hair & Rekeita, 1972; Reich, Moll & Curtis, 1976). The procedure for estimating the features for the proposed speaker recognition system is discussed in the following sections.

82.1 Data collection

We have collected speech data from ten adult male speakers. Like most of the speaker recognition systems, we have considered only cooperative speakers. That is, they do not change their speaking habits intentionally. Speech recording has been done in an ordinary office environment in two different sessions with two weeks gap between sessions. The carrier sentence selected for this study was */bha:rat hama:ra: de:ś hai/* (India is our country). Twenty five repetitions of this carrier sentence were recorded for each speaker. The speakers were not given any instructions about the manner in which they should read the sentence. Speech was digitized to 12 **bits/sample** at a sampling rate of 10 **kHz.** End points of each utterance were determined by a simple threshold logic on the amplitude of the speech signal. Duration of the utterances varied from 1.2 **sec** to 2.5 **sec.** The algorithm for pitch extraction is based on the properties of group delay functions **(Yegnanarayana** & Ramachandran, 1992).

8 3 3 Extraction of input features

We have used word boundary hypothesization algorithm for the extraction of input features from the utterances of carrier sentence. **As** discussed in Chapter 7, word boundary hypothesization algorithm is based on the phonological pattern of the constituent words, it **hypothesizes** word boundaries irrespective of the speaker. The algorithm was used on the utterances of the carrier sentence by the ten speakers. The accuracy is more than 95%. In some cases errors occur due to errors in finding the peaks of the energy contour, which result in wrong placement of boundaries. These errors can be avoided using the knowledge of the inherent durations of the syllables in the sentence.

Fig.8.2 shows the word boundaries hypothesized for the natural utterance corresponding to the carrier sentence */bha:rat hama:ra: de:ś hail* (India is our country). Thick line indicates the pitch contour of the utterance and thin line corresponds to the energy contour. Vertical bars in the energy contour indicate the peaks used to determine the mid points of the syllable nuclei. Circles marked in the pitch contour correspond to the pitch frequency value at the midpoints of the syllable nuclei. The sentence has three word boundaries and the algorithm hypothesized all of these boundaries correctly.

For speaker recognition we use pitch as well as durational features. The relative contribution of these features to the proposed speaker recognition system is discussed in Section 8.4. The features related to the pitch contour are the pitch frequency values at valleys and peaks for disyllabic and trisyllabic words and the value at the midpoint of the syllable for monosyllabic words. In the carrier sentence, the first and the second words are disyllabic and trisyllabic, respectively. The pitch frequency values at the midpoints of the initial syllable (valley) and the final syllable (peak) of these words are considered. The third and the fourth words are monosyllabic words. Therefore, the pitch frequency values at the midpoints of these syllables are selected.

The durational features used in the speaker recognition systems are durations of words and total duration of the sentence. The carrier sentence has four words and the duration of each word is measured as the distance between adjacent word boundaries measured in number of analysis frames (duration of each frame is 25.6 msec with a shift of 614 msec from the previous frame). Word boundaries are determined by the word boundary hypothesization algorithm. Thus four durational features corresponding to the four words and one feature corresponding to total duration of the utterance are used for speaker recognition together with six pitch accent features. Thus the total number of features used in the speaker recognition systems is eleven. Normalization of the features were done by dividing the input values with a constant value which is higher than the maximum value of the corresponding features available.

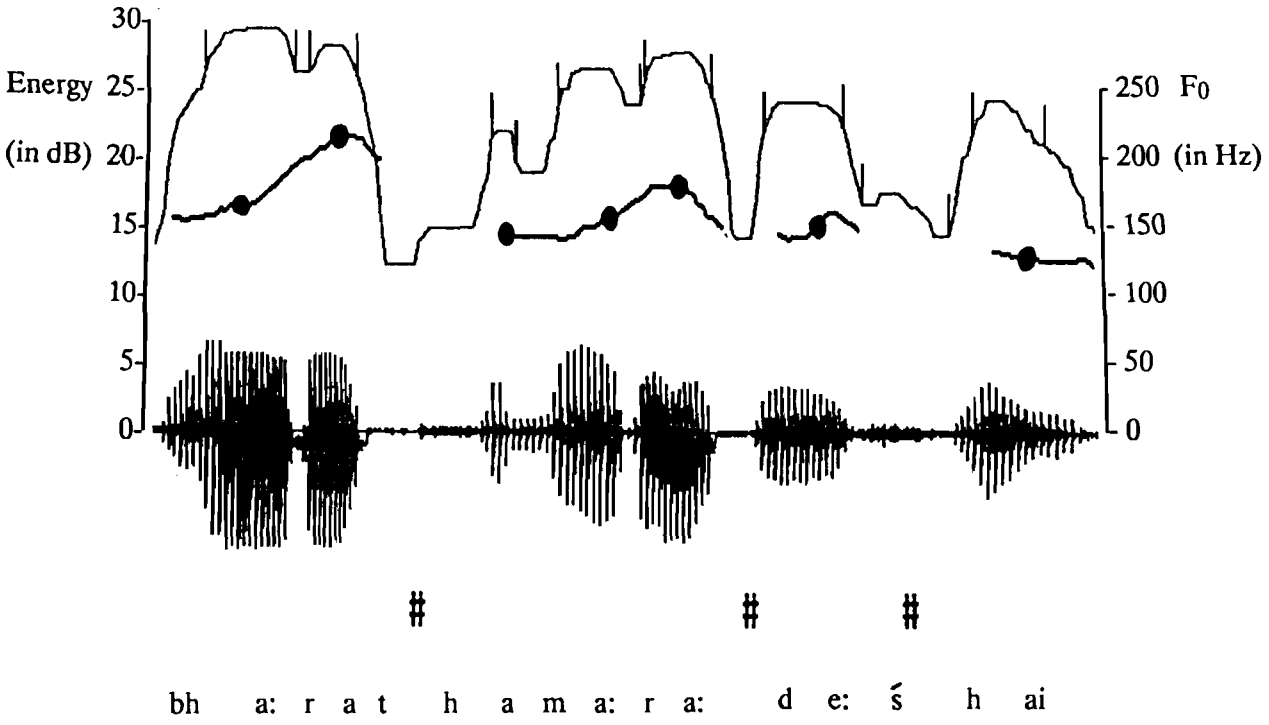


Fig.8.2. Hypothesization of word boundaries using word boundary hypothesization algorithm for the test utterance of the carrier sentence */bha:rat#hama:ra:#de:ś#hai/* (India is our country). Thick line indicates the pitch contour for the utterance and the thin line corresponds to the energy contour. Vertical bars in energy contour indicate the locations of the syllable nuclei. Circles marked in the pitch contour correspond to pitch frequency value at the midpoints of the syllable nuclei. # indicates a word boundary hypothesized by the algorithm.

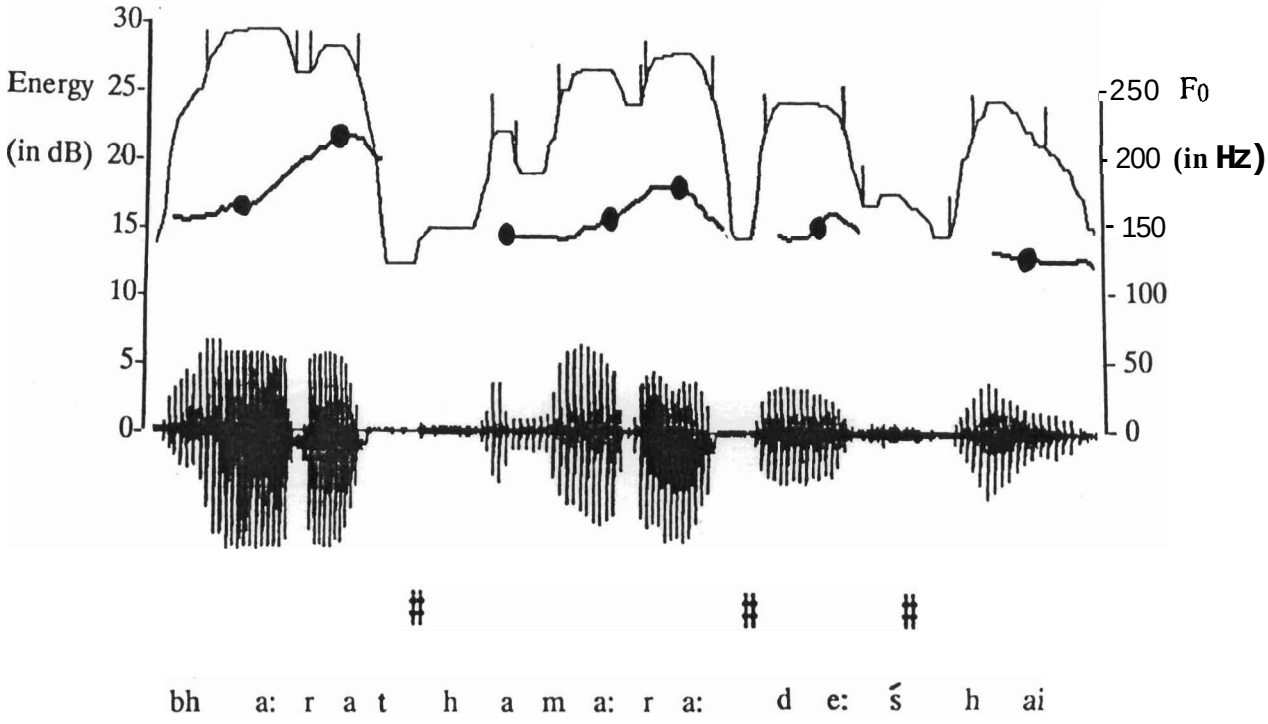


Fig.8.2. Hypothesization of word boundaries using word boundary hypothesization algorithm for the test utterance of the carrier sentence */bha:rat#hama:ra:#de:ś#hai/* (India is our country). Thick line indicates the pitch contour for the utterance and the thin line corresponds to the energy contour. Vertical bars in energy contour indicate the locations of the syllable nuclei. Circles marked in the pitch contour correspond to pitch frequency value at the midpoints of the syllable nuclei. # indicates a word boundary hypothesized by the algorithm.

8.23 Robustness of the features

The advantage of using pitch and temporal features as features for speaker recognition is their robustness even under noisy speech input conditions. But processing of speech signal to get correct pitch values under noisy input conditions is difficult. From our experiments, we have found that the pitch contour estimated using the properties of group delay functions does not change significantly due to the presence of noise. The utterances of clean speech were mixed with white noise at signal to noise levels of 20dB, 10dB, 5dB and 3dB. The pitch contour is extracted from both the clean speech and the noisy speech. Fig.8.3 shows the variations of the pitch contours when the speech signal is corrupted with various levels of white noise. Plots of pitch contours for two speakers are shown.

For hypothesizing word boundaries, the word boundary hypothesization algorithm uses the pitch accent values in the high SNR portions (syllable nuclei) of the speech signal only. So even if the pitch frequency values of the low SNR segments are affected due to the addition of noise, it does not reflect on the performance of the algorithm significantly. For example, Fig.8.4 shows the performance of the word boundary hypothesization for noisy speech (SNR 3dB). Pitch and energy contours for noisy speech are also plotted. The parameters are slightly disturbed due to the addition of noise. But the performance of the algorithm is not affected by noise, because the pitch accent pattern remained more or less the same.

8.3 The model: Neural networks for speaker recognition

Automatic speaker recognition requires a mapping between speech and speaker identity so that each possible input waveform is identified with its corresponding speaker. The system has to be trained using some reference templates of features for each speaker. The features may change slightly with respect to the change in speaking environment and time. One primary requirement of any speaker recognition system is its invariance to these minor changes in the input utterance. These requirements make artificial neural

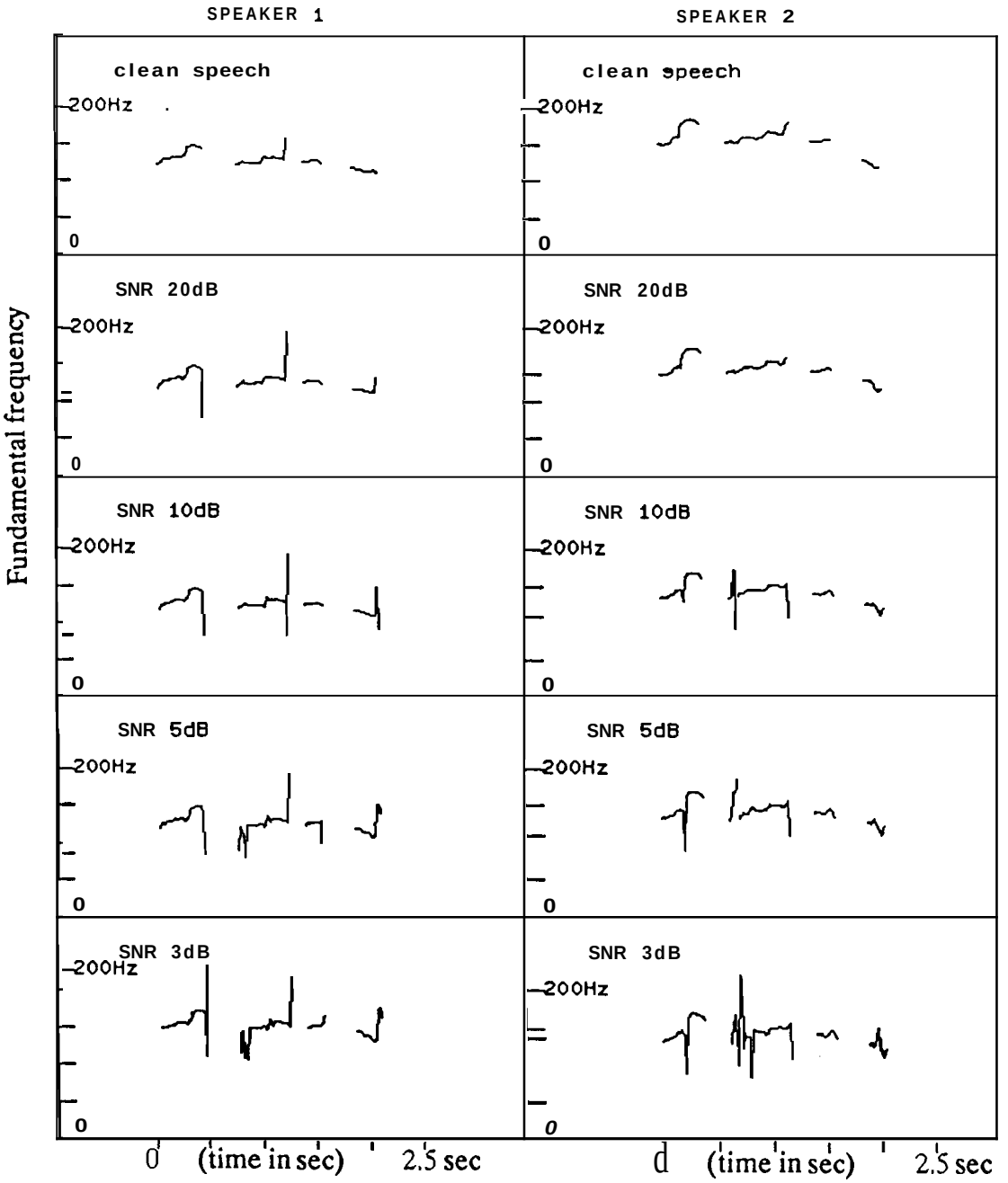


Fig.8.3. Variation of pitch contour with different noise levels. Clean speech signal is mixed with white noise at signal to noise ratio **(SNR)** levels of **20dB, 10dB,** 5dB and 3dB. Pitch contours are extracted using the properties of group delay functions. Plots of pitch contours for two speakers are given.

Fig.8.4. **Hypothesization** of word boundaries for noisy speech input conditions. The clean speech corresponding to the carrier sentence */bha:rat# hama:ra:#de:ś#hai/* (India is our country) is **mixed** with white noise **(SNR = 3dB).** Thick line indicates the pitch contour for the utterance and the thin line corresponds to the energy contour. Vertical bars in energy contour indicate the locations of the syllable nuclei. Circles marked in the pitch contour correspond to pitch frequency value at the midpoints of the syllable nuclei. # indicates a word boundary hypothesized by the algorithm.

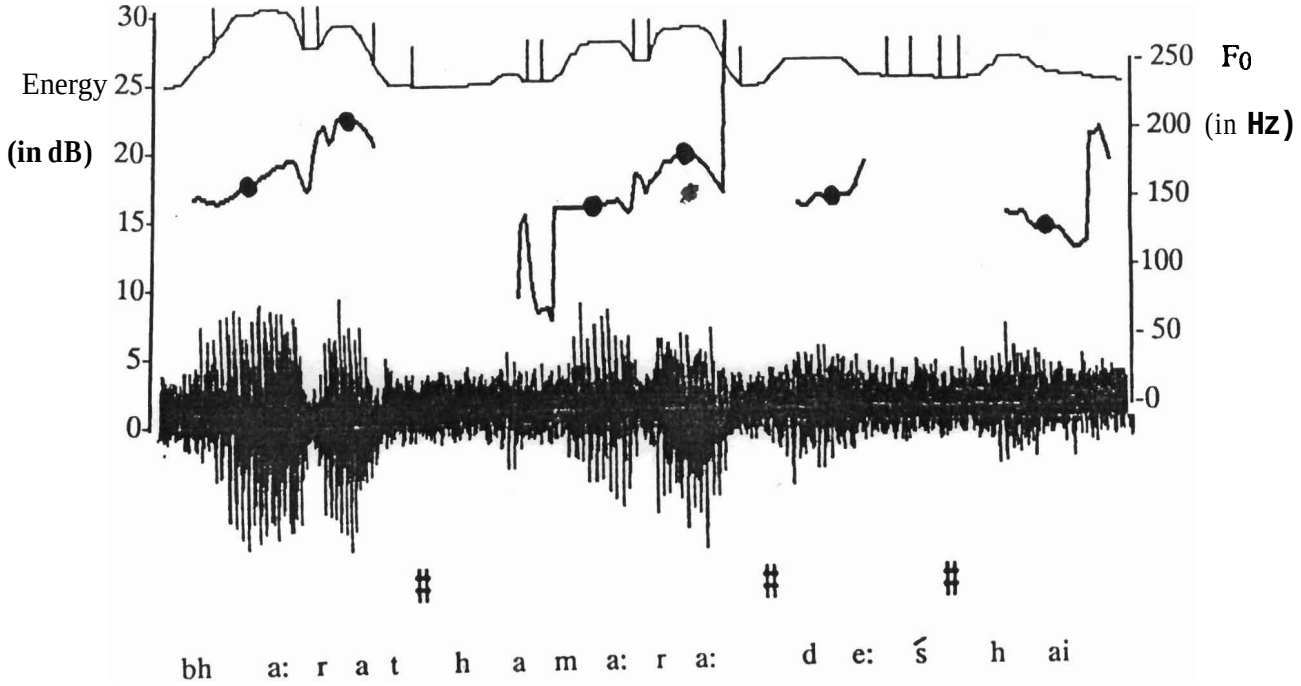


Fig.8.4. Hypothesization of word boundaries for noisy speech input conditions. The clean speech corresponding to the carrier sentence */bha:rat# hama:ra:#de:ś#hai/* (India is our country) is mixed with white noise (SNR = 3dB). Thick line indicates the pitch contour for the utterance and the thin line corresponds to the energy contour. Vertical bars in energy contour indicate the locations of the syllable nuclei. Circles marked in the pitch contour correspond to pitch frequency value at the midpoints of the syllable nuclei. # indicates a word boundary hypothesized by the algorithm.

networks a suitable choice for recognizing speakers from the input features. Artificial neural networks can modify their behavior in response to their environment. Also, they are able to generalize the input features automatically and hence it can overcome small variations due to noise and distortion.

Based on the training procedure, neural networks are classified into supervised and unsupervised categories. In supervised training, for each input vector the output of neural network is calculated and compared with the desired output in each training cycle. Weights are changed according to an algorithm to minimize the error. These neural networks are suitable for a closed set speaker identification system. That is, if the number of speakers is fixed, we can collect the reference templates for each of the speakers to train the network.

Consider the case of open set speaker identification. Here the number of speakers is not fixed and new speakers can be added at any instant. In such cases it is not practical to use a supervised training procedure. A neural network based on adaptive learning procedure can be used. The basic requirement for such a system is that it should be adaptive such that the learning of a new pattern should not erase or significantly modify the previous training.

In the following section we discuss two neural network architectures for automatic speaker recognition. They are based on back propagation algorithm and adaptive resonance theory, suitable for closed set and open set speaker identification, respectively.

83.1 Back Propagation (BP) algorithm

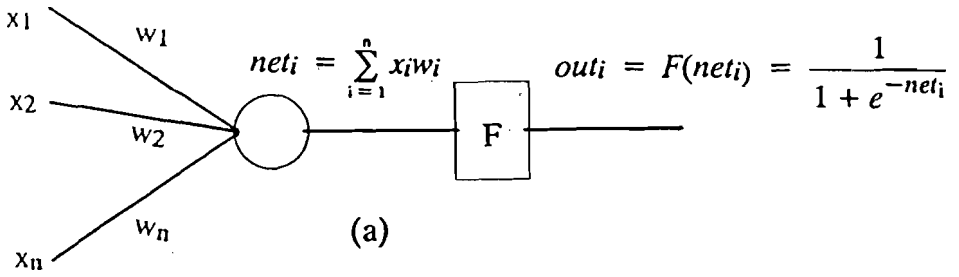
Back propagation is a systematic method for training multilayer artificial neural networks. Fig.8.5a shows an artificial neuron which is used as the fundamental building block for the network. A set of inputs {xi, $x_2, x_3,...,x_n$} is applied, either from outside or from the previous layer. Each of them is multiplied by a weight from the set $\{w_1, w_2, w_3,...,w_n\}$ and the sum of products is calculated (net_i) for each neuron i. An activation function, which is sigmoid in nature (F) is applied to modify the net input and thereby producing the output value (out_i).

Fig.8.5b shows the detailed architecture of the proposed network. Solid line in the figure indicates the forward propagation of signals and dashed line shows the backward propagation of errors. The proposed network has eleven input nodes corresponding to six pitch features and five durational features. The model is a two layer network consisting of one hidden layer and one output layer. The number of hidden nodes is selected as twenty and is selected based on a performance analysis for different number of hidden nodes. A detailed discussion on the criteria for the selection is given at the end of this section (Section 8.3.1.2). The total number of output nodes is equal to the number of speakers in the set. For the present system, it is ten. The weights are modified in order to minimize the mean squared error. This modification was done by the generalised delta rule (McClelland & Rumelhart, '1988).

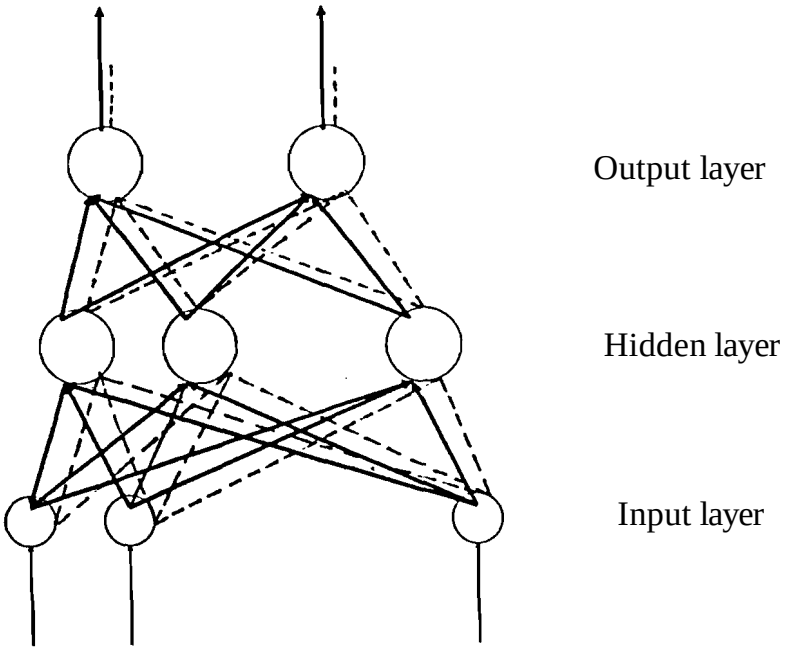
Application of the generalised delta rule involves two phases. During the first phase, the input is presented and propagated through the network to compute the activation level of each output unit. The second phase involves a backward pass through the network which is analogous to the initial forward pass. Here, the error signal is passed to each unit in the network and the appropriate weight changes are made. This allows a recursive computation of error to be propagated for each output unit. Similarly, the propagation errors for hidden units are determined recursively in terms of units to which it directly connects and the weights of these connections. These propagation errors are used to compute the weight changes in the network.

8.3.1.1 Training the network for speaker recognition

Back propagation assumes supervised mode of training. The objective of training the network is to adjust the weights so that application of a set of inputs produces the desired set of outputs. Training assumes that each input vector is paired with a target vector representing the desired output. In order to train the network for speaker recognition, we have to create a training set consisting of pitch and durational features and the code for the corresponding speaker. We created ten training pairs for each speaker. The training process can be



(a)



(b)

Fig.8.5. Architecture of a neural network based on back propagation algorithm

(a) Model of a basic neuron element.

(b) Architecture of the proposed speaker recognition system based on back propagation algorithm. The network consists of two layers, one hidden layer and an output layer. Training of this network involves two phases. Thick line indicates the forward phase of training, in which output of each neuron is computed. Dotted line corresponds to the backward propagation of errors where errors are determined as the differences between the desired and actual outputs.

summarized as follows:

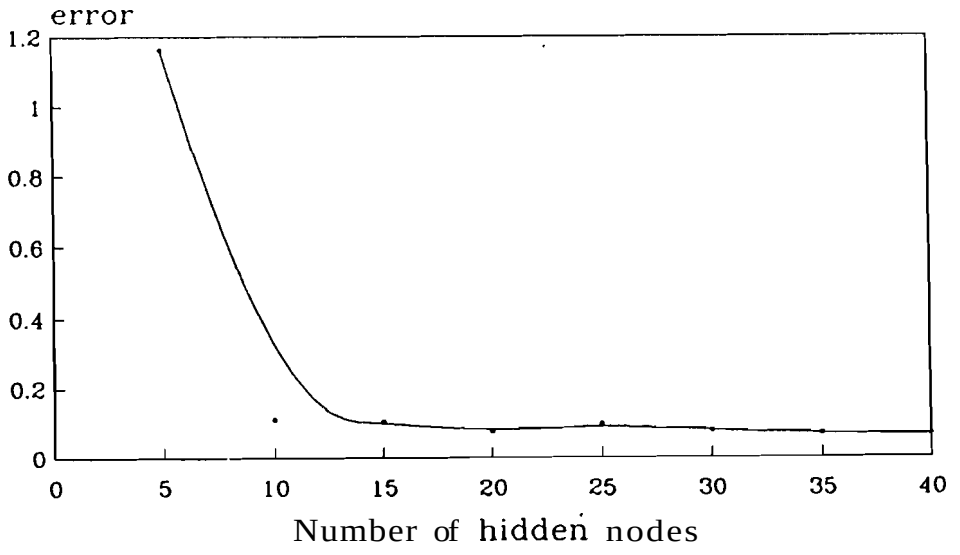
Before starting the training, all weights must be initialized to small random values. This ensures that the network is not saturated by the large values of the weights. During the training process, input vectors are given and the desired output is specified. Since we use the network as a classifier, all outputs (speakers) are set to zero except for the speaker from whom input was taken. The desired output is one. The inputs from the training set are presented cyclically until the network stabilizes. During training actual output errors are calculated and the weights are adapted based on the recursive algorithm discussed above. The algorithm stops after a specified number of training cycles (epochs), or if the errors reached within the specified limits, which ever is earlier. When the training finishes, network is considered as ready for testing and the weights can be stored in a file for future recalling.

8.3.1.2 Selection of number of hidden nodes and the number of training cycles for a network

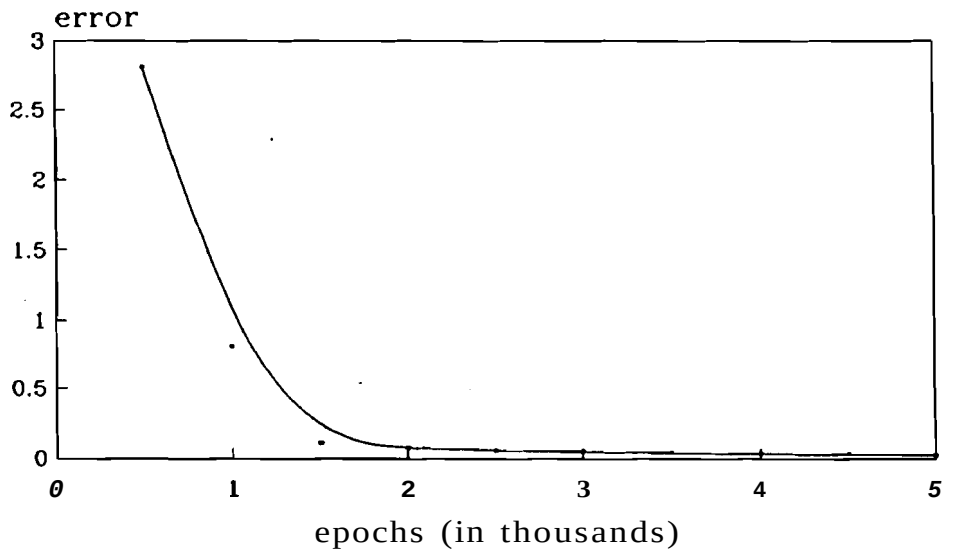
The criterion for selecting the number of hidden nodes and the number of epochs for training is based on the convergence of the network during training. Fig.8.6a shows the mean square error variation with respect to the number of hidden nodes in the network for constant number of epochs (number of epochs selected is 2000). The more the number of hidden nodes, the more the computation required. Here the error is reducing rapidly when we increase the number of hidden nodes from five to twenty. After that the error is not reducing significantly with increase in the number of hidden nodes. Hence we have selected fifteen hidden nodes. Fig.8.6b shows the relation of the number of epochs with the mean squared error. The total number of hidden nodes is taken as constant (twenty). The error is decreasing rapidly upto 2000 epochs and it is not changing significantly thereafter even if we increase the number of epochs from 2000 to 5000. Hence we have selected the number of epochs as 2000.

8.3.2 Adaptive Resonance Theory (ART)

Neural networks based on the adaptive resonance theory have several



(a)



(b)

Fig.8.6. Number of hidden nodes and the number of training cycles for a network.

(a) Variation of mean squared error with respect the number of hidden nodes. Number of epochs for the training is taken as constant (2000epochs).

(b) Variation of mean square error with respect to the number of epochs. Number of hidden nodes is taken as constant (20).

advantages. Development of this model is motivated by the function of brain which is able to receive new information as they arrive (plasticity) without changing the stability needed to ensure that the existing information are not erased or corrupted in the process. ART is ideally suitable for changing environment (Grossberg, 1987). Open set speaker identification is such a task because we can expect the addition of a new speaker at any time. ART resolves this by the addition of a classification category and declares the speaker as new without changing the existing memory. In the following section we discuss the architecture and characteristics of the ART networks.

8.3.2.1 Architecture of ART network

Adaptive resonance theory is divided into two paradigms, each divided by the form of input data and its processing. **ART1** is designed to accept only binary input vectors. ART2 can classify both continuous and binary inputs. The proposed speaker recognition system is based on **ART2.** In the following paragraphs we discuss the architecture of an ART2 network.

Fig.8.7 shows a typical ART architecture. It consists of two fields. They are: 1) a feature representation field **(F_1)** and 2) a category representation field (F2) (Carpenter & Grossberg, 1987). These fields correspond to the short term memory (STM) of human brain. Nodes of these fields undergo cooperative and competitive interactions. The connection from the field F_1 to F_2 is through a bottom-up adaptive filter which encodes new input patterns by changing weights. The connection from the field F_2 to F_1 is through a top-down adaptive filter which leads to the property of the code stabilization. These correspond to the long term memory (LTM) of the human brain and plays the role of learned expectation in an ART system. The fields F_1 and F2 together with the bottom-up and top-down adaptive filters constitute the **ART's** attentional subsystem.

ART network consists of an orienting subsystem which becomes active when a bottom-up input to F_1 fails to match the learned top-down expectation read out by the active category representation at F_2. This activation resets the field F_2 and induces the attentional subsystem to proceed with a parallel search.

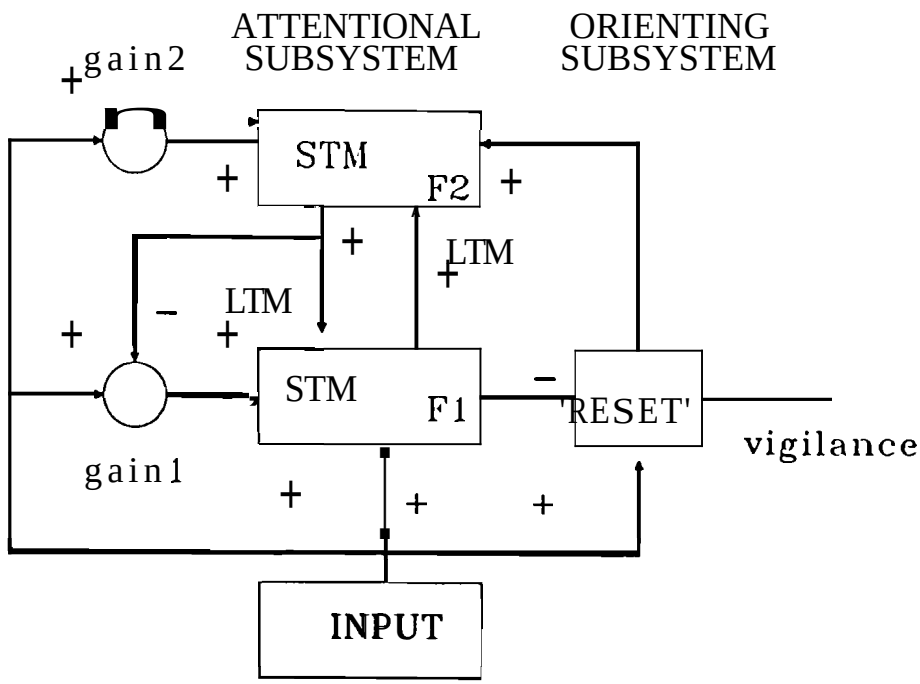


Fig.8.7. Architecture of a neural network based on adaptive resonance theory. F1 and F2 are feature representation and category representation field, respectively, analogous to the short term memory (STM) of brain. LTM, long term memory encodes new input patterns as the traces of bottom-up and top-down adaptive filter representations. These together called the attentional subsystem of the network. Gainl, gain2 and vigilance are the parameters of orienting subsystem which helps for the stabilization of the network (Carpenter & Grossberg, 1987).

This search procedure is updated adaptively through out the learning process. A category is established as a result of this search if it fails to find an adequate match and thus a matched F_1 pattern resonates within the system. The search for an adequate match for an input pattern can be adjusted by changing parameters such as vigilance, gain1 and gain2.

The processing cycle of an ART system consists of bottom-up adaptive filtering, code selection, read out of top-down learned expectation and code reset. The parts of this processing cycle correspond to the cognitive process of discovering, testing, searching, learning and recognizing the hypotheses. Due to these features of ART, this network is ideally suitable for an open set speaker identification system. Consider the instant of introducing a new speaker to the system. It searches the established categories for an adequate match. If the system finds a match, then it refines the adaptive filter representations, if necessary, to incorporate new patterns. If no matches are found, and the full coding capacity is not exhausted, a new category will be formed with the previously uncommitted adaptive filter traces encoding the fields established by the input pattern. Thus a new speaker gets added to the system.

8.3.2.2 Training the network

ART2 is trained in the unsupervised mode. Here the training set consists of input features without any target output. We selected ten distinct patterns for each speaker based on a preliminary analysis. The clusters formed by the training of these patterns are classified into ten groups and each group corresponds to a speaker. Later during testing, the speaker is identified from the unknown input pattern based on the cluster to which the pattern associates. If it fails to identify a proper match, then the speaker is categorized as a new speaker. A periodic updating of clusters allocated to each speaker increase the identification accuracy significantly. This flexibility of the network helps us to add new speakers to the identification system without affecting the existing system.

As discussed earlier, the parameters of the orienting subsystem play a significant role in the performance of the system. The matching criterion is

primarily determined by the vigilance parameter which controls the activation of the orienting subsystem. Higher vigilance imposes a stricter match criterion and hence partitions the input set into finer categories provided all other things remain unchanged. Lower vigilance tolerates greater mismatches in feature representation field and hence classification is done into coarser categories. Also, at every vigilance level the matching criterion is self scaling based on the complexity of the input patterns. In our system, we decided to impose a higher vigilance (> 0.9) which gives sufficiently good clustering. But later we need a grouping of these clusters to identify the speaker. Attentional gain control at F_1 and F_2 (gain1 and gain2, respectively) acts to adjust overall sensitivity to patterned inputs and to coordinate separate asynchronous functions of the ART subsystems. Another significant parameter which alters the recognition accuracy is the signal threshold used for defining level of variation. A proper selection of this parameter helps the system to treat similar patterns in the same way and to find out the unfamiliar patterns separately. We selected the value of this parameter as inverse of the square root of the number of input features (0.3014).

8 3 3 Comparison between back propagation and adaptive resonance theory

The domain of applications for back propagation and adaptive resonance theory are entirely different. Back propagation is suitable for a condition where the domain is clearly prespecified. That is, a domain in which it is possible for a systematic training and then recall as in the case of a closed set speaker identification. The algorithm is comparatively straight forward and easy to implement. But it fails to solve the stability-plasticity dilemma (Wasserman, 1989). This surfaces as a major disadvantage when the network is exposed to a changing environment like an open set speaker identification. The advantage of ART network is its capability of learning new patterns without modifying the previously learned patterns (Grossberg, 1987; Lippmann, 1987).

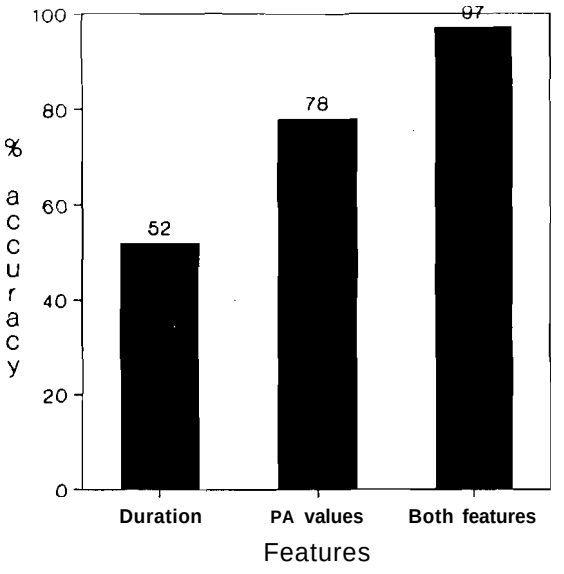
Enough training should be given for a BP model before testing it with a new input set. On the other hand ART model is self stabilized in any arbitrary input environment. ART model learns by itself without the knowledge of the output

and hence the training is unsupervised. The concept of matching is different for both the models. In ART model matching alters the information processing and regulates the learning process. Matching in **BP** model does not have any effect on the information processing. All computations of ART are in real time and there is no assumption of weight transport as in the case of back propagation.

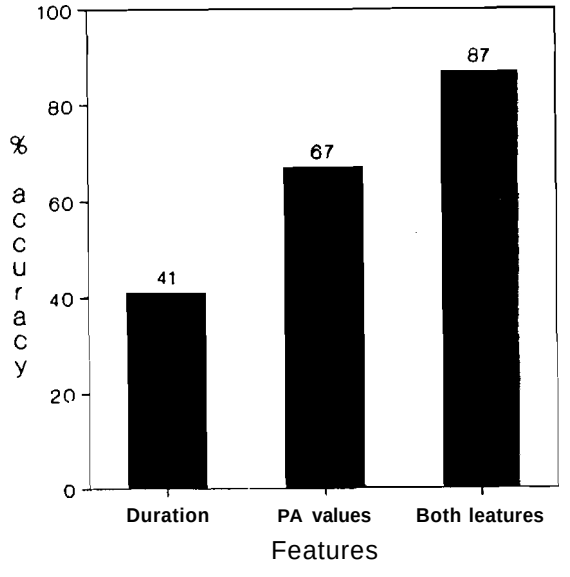
8.4 Results and discussion

As discussed earlier, the system was based on a fixed text and was tested for ten speakers. For this 25 utterances of same text was collected from each speaker. Of these, ten utterances were used for training and the remaining for testing. The performance of the system was analysed for both the clean speech and the noisy speech.

To analyze the effect of durational and pitch accent values, we perform speaker recognition using these features alone. Number of input nodes for a durationally guided speaker recognition system is five, correspond to durations of four words in the carrier sentence and the total duration of the utterance. The system is tested for clean speech. It gave an accuracy of 52% for the neural network based on **BP** algorithm and **41%** for the neural network based on ART. Number of input nodes for a speaker recognition system with pitch accent values as input features **is** six. These correspond to the pitch accent values of the initial and the final syllables for disyllabic and trisyllabic words and the pitch accent value at the midpoint of the syllable for monosyllabic words in the carrier sentence. It gave an accuracy of 78% for **BP** model and 67% for ART model. From the results, it is obvious that pitch accent values are more reliably identifying the speakers than durational values. When combine both the features, the performance of the system improves significantly. Also, it is desirable to consider more features together due to the relative easiness to mimic the voice characteristics in a system when it uses any one of the prosodic feature in isolation. The accuracy of the proposed system is discussed in the following section. Fig.8.8 shows improvement in the performance of the speaker recognition system by considering more prosodic features together.



(a)



(b)

Fig.8.8. Improvement in the performance of speaker recognition by considering pitch and durational features together.
(a) In the neural network based on back propagation algorithm
(b) In the neural network based on adaptive resonance theory

8.4.1 Results on clean speech

The accuracy for speaker recognition system was tested for both the neural network architectures. The number of hidden nodes in the network and the number of epochs for the training were decided based on the mean square error of the system as discussed in Section 8.3.1.2. Classification results of the speaker recognition system based on the back propagation algorithm are shown in the confusion matrix in the Fig.8.9a. The column X corresponds to the number of unrecognized utterances. That is, the activation levels of the output nodes were less than the threshold level (0.5) and hence none of the output nodes were selected. The overall percentage of correct speaker identification of speakers in this system is 96.25%.

Fig.8.9b shows the confusion matrix of the speaker recognition system based on ART2 algorithm. The column X corresponds to the number of unrecognized utterances. These utterances were selected as the utterances of a new speaker. The overall percentage of correct identification of speakers in this system is 86.88%. Even though the accuracy is less compared with BP algorithm, ART has the advantage that it can be used in a changing environment due to its adaptive nature.

8.42 Results on noisy speech

The advantage of designing speaker recognition systems based on intonation knowledge is the robustness against noisy speech input conditions. We tested the performance of the speaker recognition systems for noisy speech also. Pitch accent values in high SNR portions of speech are considered and hence the presence of noise does not reflect on the extraction of the features significantly.

In order to check the performance of the system in the noisy speech, we arbitrarily selected 15 utterances of each speaker and the speech signal is mixed with white noise to obtain SNR levels of 20dB, 10dB, 5dB and 3dB. The features were extracted as in the case of clean speech. The speaker recognition system was tested for both the BP and ART2 algorithms. Fig.8.10 shows the variation of

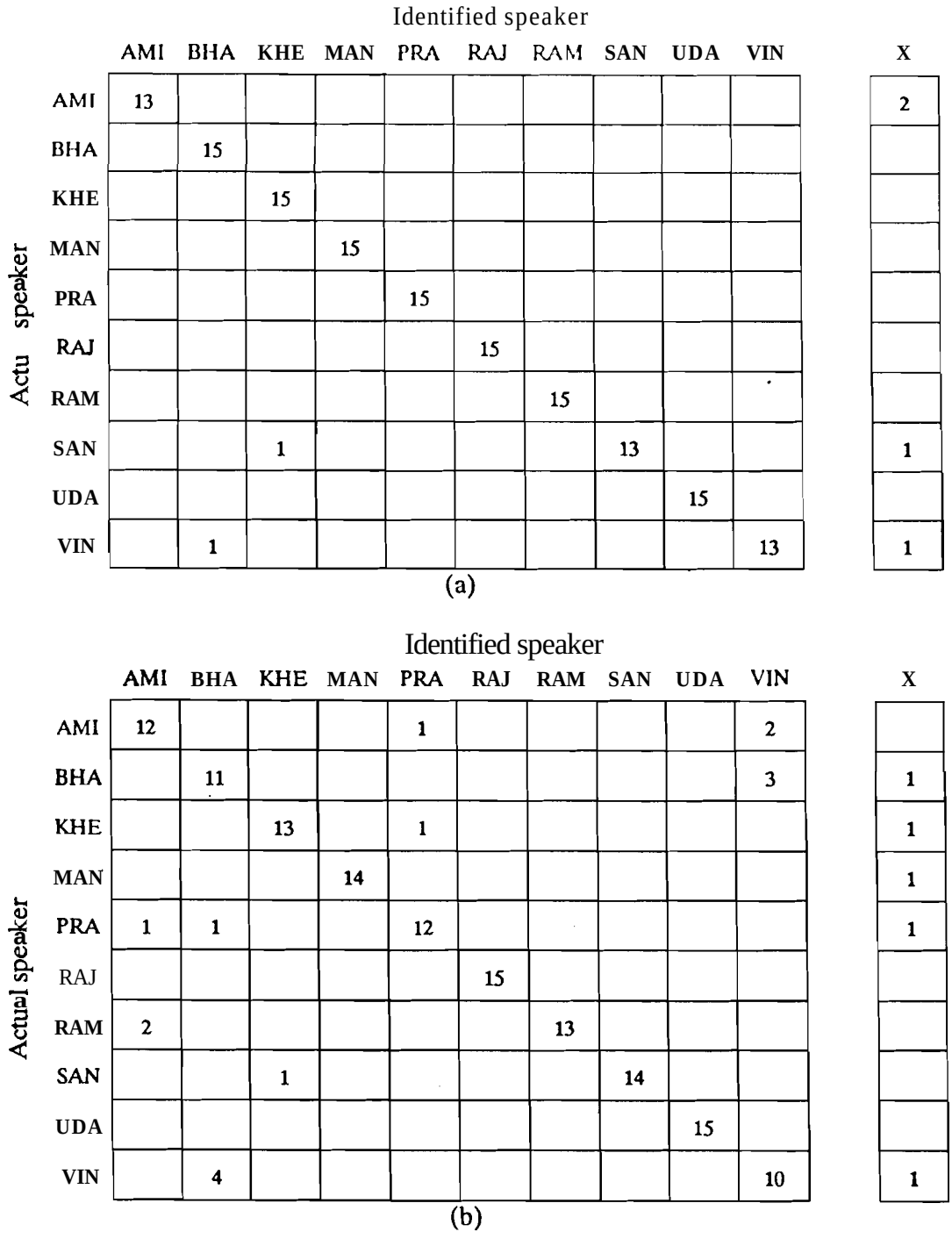


Fig.8.9. Confusion matrix obtained for the speaker recognition in clean speech.

(a) Using back propagation algorithm

(b) Using adaptive resonance theory

Rows in the matrix: indicates the actual speaker and the columns indicate the identified speaker. Column X corresponds to the unrecognized speech utterances. The accuracy of speaker recognition for clean speech using BP algorithm is 96.25% and using ART algorithm is 86.88%

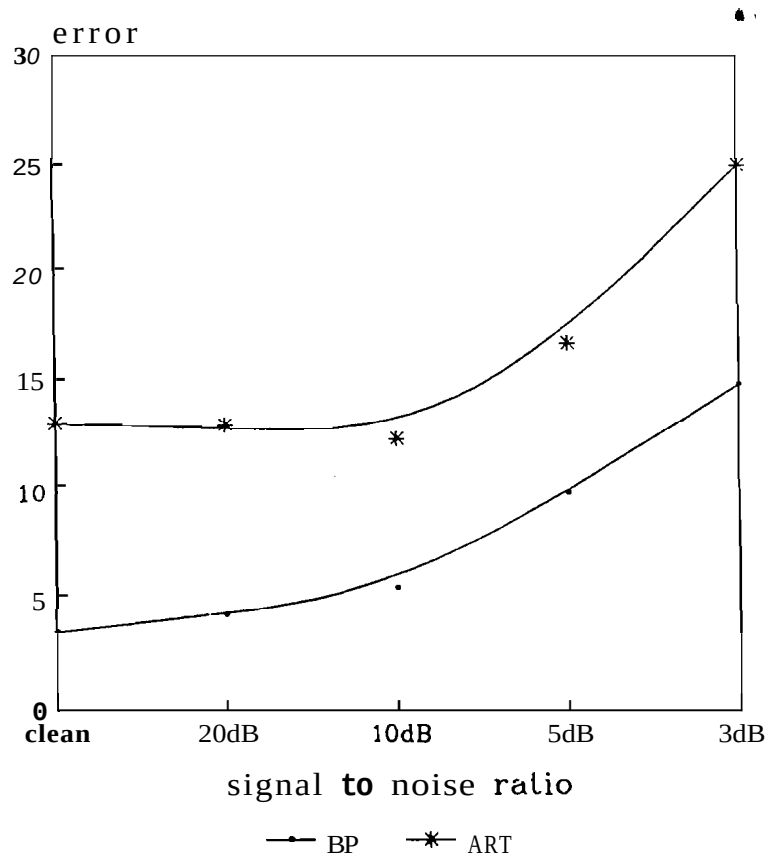


Fig.8.10. Performance degradation in the speaker recognition due to the addition of noise in speech. Error rates for a BP based system are 3.75%, 4.38%, 5.63%, 10% and 15% for clean speech, and noisy speech signals of SNR 20dB, 10dB, 5dB and 3dB, respectively. Similarly, the corresponding error rates for ART based system is 13.12%, 13.12%, 12.50%, 16.88% and 25%.

percentage errors for various noise levels for the systems. The accuracy of the BP based system for clean speech was 96.25%. It reduced to 95.63%, 94.38%, 90% and 85% for the SNR levels of 20dB, 10dB, 5dB and 3dB, respectively. Similarly for ART2 based system the accuracies are 86.88%, 86.87%, 87.5%, 83.12% and 75% for clean speech, and for the speech signals at SNR levels of 20dB, 10dB, 5dB and 3dB, respectively. The performance of both the systems was not degraded significantly due to the addition of noise upto 5dB. When the SNR reduced to 3dB, the error in the speaker identification increased significantly.

The addition of noise introduces small errors at the feature extraction stage. These are mainly due to errors in computing the energy peaks that correspond to the positions of syllable nuclei. From our experiments, we have found that the errors in the word boundary hypothesization is around 10% if the SNR of the speech signal is 3dB. The errors can be classified into three categories. They are: 1) extra boundaries, 2) missed boundaries and 3) wrongly placed boundaries. Most of the errors belong to the class of wrongly placed boundaries. These errors can be reduced significantly by using the knowledge of minimum possible duration of each word in the utterance of the carrier sentence.

8.5 Summary

A speaker recognition system based on pitch accent features has been presented. The input features for the speaker recognition system were extracted using word boundary hypothesization algorithm based on pitch accent patterns. Small errors in feature extraction can be corrected using durational knowledge because of the fixed text nature of the system. We discussed two neural network models suitable for different domains of speaker recognition tasks. Models based on back propagation algorithm for closed set identification and adaptive resonance theory for open set identification were suggested. For clean speech the former system gives an accuracy of 96.3% and the latter gives 86.9%. Since the input parameters are robust against noise, it worked well for noisy input conditions even upto an SNR of 3dB.

Chapter 9
SUMMARY AND CONCLUSIONS

9.1 Summary of the work done

In this thesis we have studied some issues in the acquisition of intonation knowledge from continuous speech and in the incorporation of the intonation knowledge in various speech systems for Hindi. In the following paragraphs we summarize our observations.

Like several other languages, the F_0 contour for declarative sentences in Hindi also exhibits a declining tendency. Properties of declination of Fo contour in Hindi can be summarized as follows: (1) Declination of F_0 contour in Hindi is characterized by falls (valleys) and rises (peaks). (2) These falls and rises fluctuate between two abstract lines -- a top line and a base line, drawn near or through all maxima and minima values of F_0 contour in a sentence, respectively. (3) The difference between a valley and the following peak (range of F_0 contour) decreases with time. (4) In a neutral declarative sentence the maximum value of F_0 will be located in the first content word itself. (5) In connected speech, the content word together with the preceding or the following monosyllabic function words form a pitch accent group called prosodic word under certain conditions.

Interrogative sentences in Hindi can be broadly classified into two. They are (1) yes-no type interrogative sentences and (2) question-word type interrogative sentences. Intonation patterns for both types of interrogative sentences are different. Questions expecting yes-no answers have a continuous rise in Fo contour and hence the top line and the base line rise towards the end. The intonation pattern for a question-word type interrogative sentence exhibits a

dual nature. The top line and the base line decline gradually up to the question word and then rise towards the end. This backdrop declination or rising is characterized by local falls and rises.

Fo contours of prosodic words in Hindi exhibit a regular pattern of a valley precedes each peak. The following are some of the general features of valleys and peaks of F_0 contour obtained by analyzing large amount of data: (1) The valleys and peaks are mostly associated with the vowels which are the nuclei of the syllables. (2) If the prosodic word is monosyllabic then the valley and the peak occur within the same syllable and hence F_0 rises steadily. (3) In the case of disyllabic and trisyllabic words peak occurs on the final syllable and the valley occurs on the initial syllable. (4) Tetrasyllabic words show two patterns: (a) a valley on the initial syllable and a peak on the final syllable and (b) the valley and the peak occur on alternate syllables and hence characterized by two valleys and two peaks. The difference between two patterns is caused by the number of morphemes in the word. (5) The pattern for pentasyllabic words is similar to a combination of disyllabic and trisyllabic words.

F_0 contour gets modified across major syntactic boundaries which is called resetting of F_0 contour. The resetting is used as a marker for phrase boundaries and it is accompanied by a significant amount of pause. Resetting of F_0 will take place both in valleys and peaks. But the magnitude of resetting differs in both cases. The *initial peak* Fo (F_0 value of the first peak of the first intonation phrase) is constant for a particular speaker. All other significant peaks and valleys in the subsequent clauses can be related to the initial peak F_0. Various physiological, syntactic and semantic constraints affect the resetting of F_0 contour significantly.

Speakers give different durations of pauses between words while uttering a continuous text. The duration of pause between words is controlled by different features like lexical content of the words, position of the word in an intonational phrase and the phonetic factors of post-pause and pre-pause syllables.

The intonation pattern of an utterance is also affected by segmental factors of constituent units. Values of F_0 for vowels were studied by embedding test words in a carrier sentence. From this analysis we have found that there is a

correlation between the height of the vowel and its inherent F_0. If other factors remain constant, high vowels exhibit high F_0 than low vowel. If the quantity of the vowel increases without changing any other factor, then inherent F_0 also increases. In all the cases final vowel have greater values of F_0 than the initial vowel. Within each test word further regular variations in F_0 were observed by changing the surrounding speech units. However, these changes are very small when compared with the changes due to other properties of F_0 contour.

Intonation knowledge acquired by the above analysis was successfully incorporated in text-to-speech, speech-to-text and speaker recognition systems. It was used as a higher level knowledge source to improve the performance of these speech systems.

There are different stages involved in the incorporation of intonation knowledge in a text-to-speech system. They are summarized as follows: (1) Input text has to be parsed to find out the type of sentence and corresponding intonation behavior. (2) Text analysis has to be performed both at the word level and at the character level to incorporate various properties of intonation patterns. (3) Pitch accent rules have to be incorporated to decide the valleys and peaks of each word. (4) Rules corresponding to the changes of Fo for each syllable due to the phonetic properties of surrounding speech units have to be incorporated. (5) Proper amount of pauses has to be incorporated between words, intonation phrases and sentences. (6) Intonation knowledge has to be represented using a suitable knowledge representation scheme in order to incorporate in a text-to-speech system. (7) Activation of intonation knowledge is achieved by means of a rule based inference engine with forward chain control strategy.

Intonation knowledge is useful for hypothesizing word boundaries from continuous speech. We have developed an algorithm based on pitch accent patterns in Hindi to hypothesize word boundaries. The word boundary hypothesization algorithm was evaluated for three native speakers of Hindi with different dialect variations. The algorithm hypothesized an average of 80% of word boundaries correctly. The performance of the algorithm was evaluated for

noisy speech as well. Since we considered the pitch values only at high SNR segments of speech such as the mid points of syllable nuclei for computing the pitch accent features, the performance of the algorithm did not degrade much. Word boundary hypothesization algorithm can be extended further for hypothesizing unaccented monosyllabic function words in Hindi.

Intonation knowledge can be used for identifying one speaker from another. We have developed a text dependent automatic speaker identification system for a small (10 speakers) population based on pitch and durational features. Both of these features can be extracted using word boundary hypothesization algorithm. The classifier consists of a neural network based architecture to identify speaker from the input features. We have proposed two neural network based architectures for the speaker identification system based on back propagation algorithm and adaptive resonance theory. The accuracy of speaker identification using adaptive resonance theory (87%) is less compared with back propagation algorithm (97%). The advantage of using pitch and temporal features as the features for speaker recognition is their robustness under noisy speech input conditions.

9.2 Major contributions of the work

The major contributions of this thesis are as follows: (1) A model is proposed for intonation patterns in Hindi based on the type and structure of sentences, the nature of words, and the inherent phonetic properties of vowels and the contextual variations. (2) A method is proposed to incorporate intonation knowledge in a text-to-speech system for Hindi to study the significance of the knowledge in the system. (3) An algorithm is proposed to hypothesize some word boundaries for continuous speech in Hindi. The algorithm also hypothesizes a few *function* words (words which have only grammatical meaning) in continuous speech. (4) A speaker recognition system for small (about 10 speakers) population is developed using some robust features of intonation of a test utterance.

9.3 Discussions on the approach

Even though we were able to characterize the properties of intonation patterns in Hindi successfully, this approach has several limitations. While developing the model of intonation, we did not consider the effects of semantic factors. But in the properties of intonation patterns for continuous speech, semantics play an important role. Also several other aspects of spontaneous speech such as influence of rate of speech, effects of emotions, etc. have not been considered. We collected the speech data in neutral phonetic context to develop a model for intonation. This approach is justified because the aim of our study is to incorporate intonation knowledge in various speech systems. Our text-to-speech system is intended to perform the function of a news reader. The speech-to-text system aims to produce the function of a dictation taking machine and the speaker recognition system is based on a fixed text. In all the cases semantic aspects played a little role.

The properties of fall-rise patterns of F_0 contour is influenced by several other factors related to morphology, coarticulation, etc.. For instance, Fo contour of a bimorphemic tetrasyllabic or pentasyllabic word consists of two valleys and two peaks determined by the position of morphemes in the word. But to analyze the number of morphemes in a word we have to conduct a detailed study on morphology which is not included within the scope of this work. Also, the influence of coarticulation behavior of speech sounds on fall-rise patterns were not studied.

In Chapter 5 we studied the inherent properties of Fo contour. For that study we selected nonsense words embedded in a carrier sentence. Results of this controlled study may not match with the results from unrestricted continuous speech (spontaneous speech). But it is very difficult to arrive at any conclusions if we perform the analysis in an unrestricted situation. Moreover, the inherent properties also depend on the position of the characters in sentences. Due to these difficulties we have concentrated on controlled studies only.

In order to incorporate intonation knowledge in various speech systems, we

have generalised the properties of intonation patterns of continuous speech in Hindi. Any generalisation results in loss of some individual components. This limitation surfaces in the incorporation of the knowledge in various speech systems. Whatever may be the number of rules developed to capture the properties of intonation patterns, it will be still inadequate because we cannot characterize the minor features accurately as exists in continuous speech. This may be partly due to the hardwired nature of representation scheme and partly due to our partial understanding of mechanism of speech production.

9.4 Directions for future studies

Properties of intonation patterns for Indian languages have not been studied systematically so far. This thesis proposes an approach to characterize the properties of intonation patterns in an Indian language. Due to similarity in phonetic behavior, this approach could be adopted to other Indian languages as well.

Several factors which alter F_0 contour of continuous speech are left out in this study. Properties of intonation patterns are changing significantly with respect to various morphologic and semantic factors. These factors have to be studied in detail. We have studied the properties of intonation patterns for declarative and interrogative sentences only. This can be extended further for other types of sentences and for various styles of speaking.

As a higher level knowledge source, intonation knowledge is not exploited to its full potential in speech systems. For instance, in the development of a speech-to-text system, intonation knowledge can be used upto the phonetic level by incorporating various segmental level properties of F_0 contour.

Intonation highlights the background information in linguistically encoded messages. Hence, it efficiently communicates linguistic information contained in a verbal message. This property may be exploited further to track a particular speaker from a group of speakers in a noisy situation. Also, the properties of intonation patterns have many potential applications in converting one person's voice to another.

APPENDIX A

Phonetic transcription of Hindi consonants and vowels

In this appendix, we list the consonants and vowels of Hindi. Phonetic transcription for all characters are given. Through out the thesis we used this phonetic transcription for representing the Hindi text.

Index	Character	Phonetic transcription
consonants:		
0		
1	क	k
2	ख	kh
3	ग	g
4	घ	gh
5	च	c
6	छ	ch
7	ज	j
8	झ	jh
9	ट	ṭ
10	ठ	ṭh
11	ड	ḍ
12	ढ	ḍh
13	ण	ṇ
14	त	t
15	थ	th
16	द	d
17	ध	dh
18	न	n
19	प	p
20	फ	ph
21	ब	b
22	भ	bh
23	म	m
24	य	y
25	र	r
26	ल	l
27	व	v
28	श	ś
29	ष	ṣ
30	स	s
31	ह	h
32	ज़	z
vowels:		
0	्	
1	अ ( )	a
2	आ (ा)	a:
3	इ (ि)	i
4	ई (ी)	i:
5	उ (ु)	u
6	ऊ (ू)	u:
7	ए (े)	e:
8	ऐ (ै)	ai
9	ओ (ो)	o:
10	औ (ौ)	au

APPENDIX B

Sample sentences for analysing the properties of intonation patterns in Hindi speech.

B1. Simple declarative sentences

1. /śobha: kamre: mẽ: baiṭhi: hai./	Sobha is sitting in the room.
2. /sudha: kita:b paḍh rahi: hai/	Sudha is reading a book
3. /laḍka: baja:r ja: raha: tha:/	The boy was going to the market.
4. /vah kutta: bhõ:k raha: tha:/	That dog was barking.
5. /ajay ne: apni: tasvi:r dekhi:/	Ajay saw his (own) picture.
6. /kita:b bahu:t mahangi: thi:/	The book was very expensive.
7. /ra:ju: ne: baḍa: dhairya hai/	Raju is very courageous.
8. /si:ta: ki: ga:y saphe:d hai/	Sita's cow is white.
9. /gupta: ke: maka:n bik gaye:/	Gupta's houses got sold.
10. /ra:m apni: bahan se: mila:/	Ram met his sister.

B2. Yes-no type interrogative sentences

1. /kya: śaṅkar pa:s hogaya: ?/	Has Sankar passed ?
2. /kya: vo: laḍka: bha:g gaya: ?/	Had that boy ran away ?
3. /kya: ka:rkha:na: khula: hui ?/	Is factory open ?
4. /kya: vah pustak acchi: hui ?/	Is that book is good ?
5. /kya: tum bombai: cale:ga: ?/	Would you like to come to Bombay ?
6. /kya: ga:ḍi: samay par hai ?/	Is train on time ?
7. /kya: a:j chuṭṭi: hui ?/	Is today holiday ?
8. /kya: uska: bukha:r utara: ?/	Has your fever come down ?
9. /kya: tum bu:ḍhe: ho: ?/	Are you old ?
10. /kya: vah duka:n band thi: ?/	Was the shop closed ?

B3. Question-word type interrogative sentences

1. /a:p kaisi: hai ?/	How do you do ?
2. /tum kahã: rahte: haĩ ?/	Where are you staying ?
3. /tumha:ra: sa:thi: kaun hai ?/	Who is your company ?
4. /ajay ke: pa:s kaun si: kita:b hai ?/	Which book is with Ajay ?
5. /a:j bho:jan mẽ: kya: bana: hai ?/	What was cooked today ?
6. /yah ka:m kab khatam ho:ga: ?/	When this work would be over ?
7. /tum yahã: kyõ: a:ye: ho: ?/	Why do you come here ?

8. *lyah sava:l kaise: hal ho:ga: ?/* How will you solve this problem ?
9. */tum ne: use: kyõ: ma:ra: ?/* Why have you beaten him ?
10. */kalkatte: me: tumne: kya: de:kha: ?/* What have you seen in Calcutta ?

B4. Relative-correlative clause of complex declarative sentences

1. *ljo: laḍka: kal a:ya: tha: vah me:re: do:st ka: bha:i: hail*
   The boy who came yesterday is my friend's brother.
2. */a:p jidhar ja: rahe: hai udhar ra:sta: khara:b hail*
   The road is bad in the way you are going.
3. *ltum ne: jitna: kahã: tha:, usse: jya:da: miṭhai: mãi le: a:ya:/*
   I brought more sweets than you asked for.
4. */ra:m ne: idhar ca:y piya:, udhar kumre: mẽ: a:ra:m kiya:/*
   Ram took tea from here and took rest in that room.
5. */mãi ne: jis laḍki: ko ga:na: sikha:ya: vah ab re:ḍiyo: par ga:ti: hail*
   The girl whom I taught music now sings for radio.
6. *lidhar pa:ni: bars raha: hai, udhar dhu:p khili: hail*
   Here it is raining where as there sun shines.
7. */yah jhi:l utni: hi: gahri: hai, jitni: vah paha:ḍi: ũ:ci: hail*
   The lake is as deep as that mountain's height.
8. */jitna: tum me:hanat karo:ge: utna: pariṇa:m pa:o:ge:/*
   The more you put effort the more you get result.
9. *ljo: bacca: so: raha: hai usko: miṭha:i: de: do:/*
   Give sweets to that child one who is sleeping.
10. */si:ta: jaisa: ca:hti: thi:, use: sundar maka:n mil gaya:/*
    Sita got a beautiful house as per her expectation.

B5. Non-relative clause of complex declarative sentences

1. */mujhe: pata: nahi: tha: ki ra:m a:ya: tha:/*
   I did not know that Ram has come.
2. */kamala: ne: kaha: hai ki use nid a: rahi: hail*
   Kamala said that she feels sleepy.
3. */sumit khuś hai kyõ:ki usko: dost a: rahe: hail*
   Sumit is happy because his friends are coming.
4. */yadya:pi ra:m bi:ma:r hai to: bhi: ka:m par a:ta: hail*
   Although Ram is ill, he still comes to work.
5. */mãi jaru:r film de:khne: ja:u:nga: ca:he: vah jo: bhi: kahe:/*
   No matter what he says I will definitely go to see the movie.
6. */a:p ne kaha: tha: isliye: usne: patr likh diya:/*
   You asked him to, that's why he wrote the letter.

7. /*ratan ko: yah laga: ki sure:; usse: na:ra:j hail*
   Ratan felt that Suresh is angry with him.
8. /*ra:jan ko: a:śa: hai ki use: naukari: mil ja:e:gi:*/
   Rajan hopes that he will get the job.
9. /*śya:m ja:nta: hai ki ra:m kahã: rahta: hail*
   Syam knows where Ram lives.
10. /*sa:t baje: tak lauṭ a:na: anyatha: to sab na:ra:j ho:nge:*/
    Come back by seven, otherwise, everyone will be angry.

B6. Compound declarative sentences

1. /*si:ma: padhne: mẽ: te:j hai aur savita: bahu:t accha: ga:ti: hail*
   Sima is smart in studies and Savitha sings very well.
2. /*re:kha: ne: kapaḍe: dhoe: aur unko: sukhne: ke: liye: ḍal diya:*/
   Rekha washed the cloths and hung them for drying.
3. /*sudha:kar bahu:t cãcal hai aur sudhi:r bahu:t śa:nt hail*
   Sudhakar is very restless and Sudhir is very calm.
4. /*ra:ju: yahã: a: sakta: hai aur me:re: ghar ṭhahar bhi: sakta: hail*
   Raju can come here and can stay at my home.
5. /*ra:je:ṣ ko: pita: ka: ta:r mila: aur vah madra:s ke: liye: cal paḍa:*/
   Rajesh received a telegram from his father and left for Madras.
6. /*rame:ṣ paḍhne: mẽ: te:j hai par me:hanat nahĩ: karta:*/
   Ramesh is smart in studies but does not work hard.
7. /*a:p c a y pi:le: ya: mãĩ a:pke: liye: ka:phi: bana: du:nga:*/
   You drink tea or I will make coffee for you.
8. /*ra:m a:ye:ga: ya: uske: pita: pho:n kare:nge:*/
   Ram will come or his father will call.
9. /*sudha:kar dhani: hai par sukhi: nahĩ:*/
   Sudhakar is rich but not happy.
10. /*suṣama: ghar mẽ: ghuśi: aur ma: ne: use: puka:ra:*/
    Sushama entered the house and her mother called her.

B7. Sentences for studying the effect of negation.

(Sentences with and without negation are given)

1. /*mãĩ ka:ṇpu:r nahĩ: ja: raha hũ:*/ I am not going to Kanpur.
   /*mãĩ ka:ṇpu:r ja: raha: hũ:*/ I am going to Kanpur.
2. /*tum bombai: nahĩ: ja:o:ge: to: pita:ji: ko: bu:ra: lage:ga:*/
   If you do not go to Bombay, father will feel bad.
   /*tum bombai: ja:o:ge: to: pita:ji: ko: bu:ra: lage:ga:*/
   If you go to Bombay, father will feel bad.

3. */a:p ne: jis laḍke: ko: bula:ya: tha:, vah a:j nahĩ: a:ya: hail*
   The boy whom you called has not come today.
   /a:p ne: jis laḍke: ko: bula:ya: tha:, vah a:j a:ya: hail
   The boy whom you called has come today.
4. */agar pa:ni: barsa: to: mâi nahĩ: a:u:nga:/* If it rains, I will not come.
   /agar pa:ni: barsa: to: mâi a:u:nga:/ If it rains, I will come.
5. *lsudha: kita:b pad kar nahĩ: ayi: hail*
   Sudha has not come after reading the book.
   lsudha: kita:b paḍ kar a:yi: hail
   Sudha has come after reading the book.
6. */agar mâi patr na: likhu: to: bhi: tum jaru:r a:na:/*
   Even if I do not write letter, you must come.
   /agar mâi patr likhu: to: bhi: tum jaru:r a:na:/
   Even if *I* write letter, you must come.
7. */a:j ka: ka:m khatm karke: mâi tumha:re: ghar nahĩ: a:u:nga:/*
   After finishing the work, I will not come to your house.
   /a:j ka: ka:m khatm kurke: mâi tumha:re: ghar a:u:nga:/
   After finishing the work, I will come to your house.
8. */din mẽ: kha:na: kha:kar mãi so:ta: nahi: hu:/*
   After taking lunch I do not sleep.
   /din mẽ: kha:na: kha:kar mãi so:ta: hu:/
   After taking lunch I sleep.
9. */śya:m Sam ko: cay nahĩ: pi:ta: hail* Syam do not take tea in the evening.
   /śya:m Sam ko: ca:y pi:ta: hail Syam take tea in the evening.
10. *ltum agar kal nahĩ: a:ye: to: kitab nahĩ: mile:gi:/*
    If you do not come tomarrow, you will not get the book.
    ltum agar kal *a:ye: to: kita:b nahĩ: mile:gi:/*
    If you come tomarrow, you will not get the book.
    /tum agar kal *a:ye: to: kita:b mile:gi:/*
    If you come tomarrow, you will get the book.

B8. Sentences for studying the effect of numerals

(Sentences with numerals, their expanded text and the corresponding meaning are given.)

1. *lye: kita:b 45 rupaye: ki: hail*
   lye: kita:b painta:lis rupaye: ki: hail
   This book costs forty five rupees.
2. */bha:rat ko: svatantrata: 15 agast 1947 ko: mili:/*
   /bha:rat ko: svatantrata: pantrah agast unni:s sau sainta:lis ko: mili:/

India got independence on fifteenth August nineteen hundred and forty seven.

3. */bha:rat ka: pra:ci:n itiha:s 5000 sa:l ka: hail*
   /bha:rat ka: pra:ci:n itiha:s panc haja:r sa:l ka: hail
   India's ancient history is of five thousand years old.
4. */ra:m ki: mo:ṭar sa:i:kil 23618 rupaye: ki: hail*
   /ra:m ki: mo:ṭar sa:i:kil te:is haja:r che: sau atha:rah rupaye: ki: hail
   Ram's motor cycle costs twenty three thousand six hundred and eighteen rupees.
5. */kursi: jo: tum khari:d kar la:ye: ho: vo: 345 rupaye: 65 paise: ki: hail*
   /kursi: jo: tum khari:d kar la:ye: ho: vo: ti:n sau painta:lis rupaye: painsaṭh paise: ki: hail
   The chair which you have purchased costs three hundred rupees and sixty five paise.
6. */praka:ś ka: ve:g lagbhag 186000 mi:l prati se:kant hail*
   /praka:s ka: ve:g lagbhag e:k lakh che:a:ssi: haja:r mi:l prati se:kant hail
   Speed of light is approximately one lakh eight thousand miles per second.
7. */is samay viśv ki: jan samkhya: 550 karo:ḍ hail*
   /is samay viśv ki: jan samkhya: sa:ḍe panc sau karo:ḍ hail
   Now the population of the world is five hundred and fifty crores.
8. */bha:rat ka: samudr tar lagbhag 3000 kilo:mi:ṭar lamba: hail*
   /bha:rat ka: samudr tar lagbhag ti:n haja:r kilo:mi:ṭar lamba: hail
   Sea coast of India is approximately three thousand kilo meter long.
9. */30 janavari: 1948 ko: maha:tma: ga:ndhi: de:ś ki: se:va: mẽ: ṣahi:d ho: gaye:/*
   /ti:s janavari: unni:s sau athta:lis ko: maha:tma: ga:ndhi: de:ś ki: se:va: mẽ: ṣahi:d ho: gaye:/
   On thirty January nineteen forty eight Mahatma Gandhi became martyr for the service of the country.
10. *lprathvi: su:rya ke: ca:ro: o:r 365 din mẽ: cakkar laga:ti:hai/*
    lprathvi: su:rya ke: ca:ro: o:r ti:n sau paimsaṭh din mẽ: cakkar laga:ti:hai/
    The earth revolves around the sun in three hundred and sixty five days.

APPENDIX C

Tabular results of data analysis

C1. Prediction of intermediate peaks for simple declarative sentences

Speaker 1

Sentence	Peak1			Peak2		
	Actual	Predicted	%Error	Actual	Predicted	%Error
1	156.36	158.06	1.09	137.72	140.77	2.21
2	149.39	155.08	3.81	145.56	144.53	0.71
3	154.25	158.18	2.55	130.39	135.55	3.96
4	152.36	153.71	0.89	133.96	136.26	1.72
5	143.45	152.92	6.60	123.39	127.89	3.65
6	152.42	151.36	0.70	135.45	129.35	4.50
7	145.36	151.81	4.44	137.89	139.84	1.41
8	154.72	160.15	3.51	150.80	153.17	1.57
9	149.89	153.20	2.21	134.26	134.81	0.41
10	160.83	165.88	3.14	154.32	158.32	2.59

C2. Prediction of intermediate peaks for simple declarative sentences

Speaker 2

Sentence	Peak1			Peak2		
	Actual	Predicted	%Error	Actual	Predicted	%Error
1	172.38	173.38	0.58	158.72	156.91	1.14
2	169.46	167.66	1.06	153.82	152.65	0.76
3	176.84	180.56	2.10	153.82	156.60	1.81
4	170.82	172.56	1.02	155.24	157.83	1.67
5	164.32	167.69	2.05	149.73	156.07	4.23
6	174.89	171.58	1.89	143.24	148.74	3.84
7	180.88	182.02	0.63	164.26	157.99	3.82
8	164.25	171.08	4.16	149.19	157.25	5.40
9	170.48	179.64	5.37	148.36	154.88	4.39
10	174.25	181.18	3.98	163.86	163.69	0.10

C3. Prediction of intermediate valleys for simple declarative sentences

Speaker 1

Sentence	Valley1			Valley2		
	Actual	Predicted	%Error	Actual	Predicted	%Error
1	125.34	127.96	2.09	110.83	114.65	3.45
2	123.78	128.29	3.64	125.79	126.34	0.44
3	120.86	124.55	3.05	114.10	116.93	2.48
4	125.68	124.32	1.08	114.64	116.58	1.69
5	129.84	126.84	2.31	115.46	113.11	2.04
6	120.32	118.02	1.91	115.78	113.97	1.56
7	123.32	126.88	2.89	122.89	118.21	3.81
8	124.32	120.20	3.31	115.82	116.70	0.76
9	130.82	132.95	1.63	125.86	122.19	2.92
10	134.82	135.44	0.46	129.67	131.85	1.68

C4. Prediction of intermediate valleys for simple declarative sentences

Speaker 2

Sentence	Valley1			Valley2		
	Actual	Predicted	%Error	Actual	Predicted	%Error
1	145.92	145.68	0.16	139.63	140.89	0.90
2	138.42	135.54	2.08	133.69	130.53	2.36
3	146.92	147.48	0.38	139.38	138.29	0.78
4	133.79	137.37	2.68	130.94	129.03	1.46
5	140.83	142.71	1.33	132.46	134.97	1.89
6	132.46	133.40	0.71	128.81	127.27	1.20
7	145.42	147.94	1.73	139.48	139.37	0.08
8	129.79	130.82	0.79	120.87	125.40	3.75
9	140.84	137.83	2.14	125.89	129.82	3.12
10	142.63	146.64	2.81	135.48	140.29	3.55

C5. Prediction of intermediate peaks for yes/no type interrogative sentences

Speaker 1

Sentence	Peak1			Peak2		
	Actual	Predicted	%Error	Actual	Predicted	%Error
1	162.56	159.26	2.03	178.24	174.33	2.19
2	156.82	158.90	1.33	164.56	171.83	4.42
3	176.39	173.22	1.80	198.45	196.62	0.92
4	164.38	168.78	2.68	184.25	183.10	0.62
5	177.94	177.38	0.31	192.86	187.53	2.76
6	178.12	170.84	4.09	198.84	192.91	2.98
7	169.46	166.90	1.51	193.45	190.34	1.61
8	186.29	179.30	3.75	195.46	191.82	1.86
9	176.98	168.22	4.95	194.86	188.08	3.48
10	184.24	181.63	1.42	196.83	193.17	1.86

C6. Prediction of intermediate peaks for yes/no type interrogative sentences

Speaker 2

Sentence	Peak1			Peak2		
	Actual	Predicted	%Error	Actual	Predicted	%Error
1	189.34	185.32	2.12	212.46	202.77	4.56
2	178.96	181.94'	1.67	199.49	195.73	1.88
3	194.25	193.53	0.37	210.86	210.90	0.02
4	195.46	195.61	0.08	230.48	219.68	4.69
5	186.28	192.34	3.25	212.86	205.89	3.27
6	184.26	181.96	1.25	205.89	199.60	3.06
7	189.43	180.99	4.46	200.26	195.94	2.16
8	192.56	189.60	1.54	215.48	209.42	2.81
9	196.34	194.01	1.19	221.28	209.73	5.22
10	192.62	184.49	4.22	212.86	202.37	4.93

C7. Prediction of intermediate valleys for yes/no type interrogative sentences

Speaker 1

Sentence	Valley1			Valley2		
	Actual	Predicted	%Error	Actual	Predicted	%Error
1	134.28	138.71	3.30	142.86	145.01	1.50
2	140.92	143.94	2.14	145.56	149.37	2.62
3	154.84	155.11	0.17	162.86	169.63	4.16
4	144.86	151.77	4.77	152.48	158.27	3.80
5	153.19	148.35	3.16	162.69	160.00	1.65
6	151.86	159.24	4.86	164.29	168.85	2.78
7	150.28	156.06	3.85	159.89	165.49	3.50
8	158.74	158.58	0.10	169.37	165.72	2.16
9	150.86	159.01	5.40	166.62	171.52	2.94
10	159.89	159.57	0.20	172.64	173.25	0.35

C8. Prediction of intermediate valleys for yeslno type interrogative sentences

Speaker 2

Sentence	Valley1			Valley2		
	Actual	Predicted	%Error	Actual	Predicted	%Error
1	159.45	159.78	0.21	174.84	174.08	0.43
2	162.12	157.20	3.03	174.36	173.38	0.56
3	171.46	177.94	3.78	180.28	189.74	5.25
4	170.28	178.71	4.95	190.78	200.29	4.98
5	160.78	167.74	4.33	172.63	177.19	2.64
6	163.48	171.76	5.06	182.48	186.72	2.32
7	160.82	164.78	2.46	169.43	179.05	5.68
8	165.28	166.86	0.96	173.64	174.91	0.73
9	175.28	180.81	3.15	190.46	201.08	5.58
10	169.87	177.33	4.39	184.96	191.41	3.49

C9. Range of F_0 for simple declarative sentence

Speaker 1

Sentence	Word1	Word2	Word3	Word4
1	40.72	31.02	26.89	22.23
2	30.09	25.61	19.77	16.89
3	42.92	33.39	16.29	9.96
4	40.98	26.68	19.32	14.01
5	39.33	13.61	7.93	10.03
6	43.10	32.10	23.67	14.62
7	49.04	22.04	15.00	11.25
8	46.89	30.40	34.98	28.50
9	29.64	19.07	8.40	7.86
10	39.20	26.01	24.65	18.84
Mean	40.19	25.99	19.69	15.42
SD	5.96	5.95	7.91	6.06

C10. Range of F_0 for simple declarative sentence

Speaker 2

Sentence	Word1	Word2	Word3	Word4
1	33.89	26.46	19.09	10.84
2	42.94	31.04	20.13	14.21
3	38.50	29.92	14.44	14.06
4	44.51	37.03	24.30	19.48
5	35.30	23.49	17.27	13.94
6	51.46	42.43	14.43	9.59
7	44.86	35.46	24.78	14.51
8	50.37	34.46	28.32	18.00
9	46.19	29.64	22.47	18.30
10	47.86	31.62	28.38	17.35
Mean	43.59	32.16	21.36	15.03
SD	5.70	5.16	4.88	3.08

C11. Range of F_0 for yes/no type interrogative sentence

Speaker 1

Sentence	Word1	Word2	Word3	Word4
1	9.58	28.28	35.38	42.47
2	9.88	15.90	20.00	32.10
3	18.86	21.55	35.59	15.87
4	16.58	19.52	31.77	16.72
5	8.25	24.75	30.17	30.98
6	9.15	26.26	34.55	7.32
7	11.21	19.18	33.56	11.53
8	11.82	27.55	26.09	25.95
9	8.56	26.12	28.24	10.08
10	9.63	25.35	24.19	11.16
Mean	11.35	23.45	29.95	20.42
SD	3.38	3.93	4.99	11.15

C12. Range of F_0 for yes/no type interrogative sentence

Speaker 2

Sentence	Word1	Word2	Word3	Word4
1	11.98	29.89	37.62	40.08
2	16.59	16.84	25.13	34.71
3	15.88	23.79	30.58	11.32
4	21.92	25.18	39.70	10.31
5	10.10	25.50	40.23	43.92
6	7.82	20.78	23.41	15.26
7	14.22	28.61	30.83	8.35
8	14.66	30.28	38.84	41.50
9	15.72	21.06	30.82	10.19
10	7.02	22.75	27.90	12.97
Mean	13.59	24.47	32.51	22.86
SD	4.25	4.11	5.88	14.30

C13. Resetting values of F_0 peaks at syntactic boundaries of relative claus of complex declarative sentences

Sentence	%dur1	%pause	$\%F_{01}$	$\%F_{0r}$	$\%F_{0f}$	%ta
1	47.23	12.01	75.42	92.25	67.27	57.85
2	45.32	11.54	72.69	89.34	68.23	53.67
3	39.26	13.51	68.34	93.17	72.14	57.20
4	53.32	11.20	69.27	91.53	65.25	55.68
5	55.78	12.54	75.14	89.43	71.15	54.82
6	59.26	10.92	70.17	87.66	67.25	52.14
7	47.81	11.99	72.15	88.72	69.89	57.28
8	42.14	13.27	69.29	94.67	71.89	56.19
9	35.26	12.45	71.78	91.28	72.25	57.02
10	39.27	13.25	71.54	90.23	69.50	56.43
Mean	46.47	12.27	71.58	90.83	69.48	55.83
SD	7.41	0.85	2.28	2.05	2.30	1.71

(%dur1 and %pause are ratios of the duration of the first syntactic clause and the duration of the pause between syntactic clauses with respect to the total duration of the sentence, respectively. $\%F_{01}$, $\%F_{0r}$, $\%F_{0f}$ and %ta are the ratios of the final peak frequency of the first syntactic clause, Initial peak frequency of the second syntactic clause (resetting frequency) , the final peak frequency of the second syntactic clause and end tapering frequency with respect to the initial initial peak frequency of the first syntactic clause, respectively.)

C14. Resetting values of F_0 valleys at syntactic boundaries of relative class of complex declarative sentences

Sentence	%iV	$\%V_{0s}$	$\%V_{01}$	$\%V_{0r}$	$\%V_{0f}$
1	69.32	77.82	69.25	74.34	59.27
2	62.15	79.13	63.32	70.15	60.14
3	70.56	75.62	62.15	67.27	60.15
4	73.65	73.66	63.67	69.18	64.43
5	67.12	72.98	60.15	67.80	65.00
6	72.23	75.25	65.92	70.14	63.42
7	70.15	70.29	62.12	69.18	63.34
8	69.27	77.56	61.20	65.23	61.34
9	70.56	72.16	60.56	67.89	62.89
10	72.15	79.27	62.34	66.90	60.04
Mean	69.72	75.37	63.07	68.81	62.00
SD	3.06	2.91	2.60	2.35	1.95

(%iV, $\%V_{0s}$, $\%V_{01}$, $\%V_{0r}$ and $\%V_{0f}$ are the ratios of the first, second and the final valley frequencies of the first syntactic clause and the first and the final valley frequencies of the second syntactic clause with respect to the Initial peak frequency of the first syntactic clause, respectively.)

C15. Resetting values of F_0 peaks at syntactic boundaries of non-relative class of complex declarative sentences

Sentence	%dur1	%pause	$\%F_{01}$	$\%F_{0r}$	$\%F_{0f}$	%ta
1	44.23	9.98	73.35	87.78	67.43	55.76
2	39.29	12.67	72.18	85.12	67.65	54.89
3	45.27	9.71	67.24	86.34	70.12	56.34
4	44.17	11.87	71.57	92.88	66.26	57.13
5	53.28	10.75	67.20	88.31	72.12	55.94
6	57.98	9.56	71.23	86.56	65.56	53.25
7	38.26	11.76	68.92	87.65	68.45	55.16
8	33.14	12.01	71.52	93.95	70.98	55.29
9	55.45	10.34	67.93	88.83	68.78	54.16
10	56.87	10.19	68.45	82.81	64.67	56.92
Mean	46.79	10.88	69.96	88.02	68.20	55.48
SD	8.22	1.05	2.13	3.17	2.26	1.14

("Wurl and %pause are ratios of the duration of the first syntactic clause and the duration of the pause between syntactic clauses with respect to the total duration of the sentence, respectively. $\%F_{01}$, $\%F_{0r}$, $\%F_{0f}$ and %ta are the ratios of the final peak frequency of the first syntactic clause, initial peak frequency of the second syntactic clause (resetting frequency) . the final peak frequency of the second syntactic clause and end tapering frequency with respect to the initial Initial peak frequency of the first syntactic clause, respectively.)

C16. Resetting values of F_0 valleys at syntactic boundaries of non-relative class of complex declarative sentences

Sentence	%iV	$\%V_{0s}$	$\%V_{01}$	$\%V_{0r}$	$\%V_{0f}$
1	70.56	75.65	60.34	72.12	59.68
2	71.12	76.54	65.43	71.15	61.34
3	67.76	77.34	61.22	69.34	57.25
4	72.17	72.53	60.78	71.78	63.80
5	69.98	74.35	61.26	66.43	59.43
6	70.67	75.89	62.24	68.25	59.28
7	68.34	69.34	60.32	66.24	56.34
8	69.89	75.14	61.56	64.90	60.98
9	72.32	70.28	60.98	65.88	57.35
10	71.55	72.23	59.23	63.26	56.46
Mean	70.44	73.93	61.34	67.94	59.19
SD	1.43	2.57	1.57	2.92	2.28

(%iV, $\%V_{0s}$, $\%V_{01}$, $\%V_{0r}$ and $\%V_{0f}$ are the ratios of the first, second and the final valley frequencies of the first syntactic dause and the first and the final valley frequencies of the second syntactic clause with respect to the initial peak frequency of the first syntactic clause, respectively.)

C17. Resetting values of F_0 peaks at syntactic boundaries of compound declarative sentences

Sentence	%dur1	%pause	$\%F_{01}$	$\%F_{0r}$	$\%F_{0f}$	%ta
1	43.73	17.22	82.66	96.82	65.32	53.76
2	46.74	13.52	81.76	93.24	70.26	55.23
3	37.21	14.57	80.21	92.16	71.25	57.90
4	49.54	14.86	75.29	90.88	71.18	54.41
5	52.34	13.01	77.36	91.32	70.45	58.97
6	57.12	13.67	78.89	90.23	68.34	55.82
7	44.46	14.52	83.28	90.62	70.47	55.13
8	36.18	15.33	70.18	94.75	68.80	54.32
9	45.51	12.58	80.16	95.38	74.40	58.97
10	42.46	13.08	80.26	95.42	69.63	57.59
Mean	45.53	14.24	79.00	93.08	70.01	56.21
SD	6.07	1.31	3.71	2.25	2.21	1.87

(%dur1 and %pause are ratios of the duration of the first syntactic clause and the duration of the pause between syntactic clauses with respect to the total duration of the sentence, respectively. $\%F_{01}$, $\%F_{0r}$, $\%F_{0f}$ and %ta are the ratios of the final peak frequency of the first syntactic clause, initiil peak frequency of the second syntactic clause (resetting frequency), the final peak frequency of the second syntactic dause and end tapering frequency with respect to the initial initial peak frequency of the first syntactic clause, respectively.)

C18. Resetting values of F_0 valleys at syntactic boundaries of compound declarative sentences

Sentence	%V	$\%V_{0s}$	$\%V_{01}$	$\%V_{0r}$	$\%V_{0f}$
1	69.52	80.42	70.78	74.23	60.14
2	71.43	80.47	65.01	72.68	65.25
3	67.69	80.45	64.53	73.46	69.23
4	72.45	81.23	68.13	77.23	68.54
5	71.67	82.54	69.45	73.24	66.35
6	69.27	80.02	65.36	75.27	65.78
7	70.03	79.26	64.00	72.13	68.12
8	72.23	83.57	68.54	75.26	60.24
9	73.21	80.28	65.13	77.12	70.56
10	74.26	82.15	70.12	72.14	66.23
Mean	71.18	81.04	67.11	74.28	66.04
SD	1.91	1.26	2.43	1.80	3.32

(%V, $\%V_{0s}$, $\%V_{01}$, $\%V_{0r}$ and $\%V_{0f}$ are the ratios of the first, second and the final valley frequencies of the first syntactic clause and the first and the final valley frequencies of the second syntactic clause with respect to the initial peak frequency of the first syntactic clause, respectively.)

C19. Effect of preceding consonant on present vowel

Conso -nants	Vowels									
	a	a:	i	ɪ:	u	u:	e:	o:	ai	qu
k	148	152	169	173	170	179	160	165	153	162
c	150	152	166	173	173	179	166	165	150	159
t	148	150	166	173	173	176	163	165	153	162
t	148	150	171	176	167	179	160	165	153	159
p	150	152	166	171	176	179	163	172	153	159
kh	148	154	167	174	176	186	163	165	150	162
ch	151	160	168	178	172	179	163	165	150	166
th	150	152	171	174	176	183	163	172	150	156
th	150	152	171	178	183	183	163	168	150	159
ph	148	156	168	178	172	179	166	162	151	162
g	144	148	161	166	167	176	160	159	143	153
j	141	150	161	163	165	173	151	159	143	150
ḍ	138	150	159	166	160	170	157	162	144	156
ḍ	139	150	159	166	160	173	160	162	143	150
b	141	148	161	163	167	173	154	159	147	153
gh	146	144	159	169	167	176	157	165	144	153
jh	146	146	159	171	162	173	160	162	147	153
dh	148	146	164	168	162	176	157	162	144	156
dh	147	150	163	169	165	176	157	159	141	150
bh	142	148	164	171	167	176	157	159	147	150
n	144	144	161	163	162	173	154	159	141	153
m	142	144	161	169	167	170	157	162	144	153
r	142	144	159	166	162	170	160	162	144	153
l	144	144	163	169	165	173	157	159	143	150
ś	148	148	171	178	176	179	163	172	153	159
ṣ	150	154	169	178	176	190	166	172	153	169
s	150	150	176	178	173	183	166	169	147	156
h	150	150	171	176	173	183	163	168	150	159
v	142	146	159	169	165	170	160	162	141	153
ɣ	144	144	161	166	162	173	157	162	144	150
avg	145	149	164	171	168	176	160	163	147	156

C20. Effect of following consonant on present vowel

Conso -nants	Vowels									
	a	a:	i	i:	u	u:	e:	o:	ai	ou
k	116	122	134	146	137	149	131	132	121	128
c	119	124	135	148	139	150	132	133	122	129
ṭ	116	122	133	143	136	145	129	132	120	125
t	115	'120	129	141	134	145	127	131	118	125
p	115	120	129	141	132	141	126	129	117	123
kh	116	124	133	144	136	149	130	131	118	125
ch	120	126	136	147	139	151	134	132	122	129
th	115	123	132	144	136	147	128	131	118	124
th	115	122	130	143	134	144	128	127	119	123
ph	114	120	130	140	133	142	127	126	116	123
g	116	123	133	144	136	147	130	132	121	125
j	118	125	135	147	138	153	132	131	121	128
ḍ	114	123	133	144	136	147	129	130	118	126
d	116	122	131	141	133	143	128	127	118	122
b	115	120	129	138	132	142	128	128	115	122
gh	117	122	133	145	136	148	132	131	120	125
jh	119	126	136	148	140	149	131	133	122	127
ḍh	116	123	131	142	135	147	128	131	118	124
dh	115	121	130	141	133	144	127	129	117	122
bh	115	119	128	138	133	142	127	128	114	122
n	115	121	129	140	133	143	127	129	119	122
m	114	119	128	138	132	141	126	130	118	122
r	120	127	138	149	140	153	134	137	122	129
l	119	125	137	148	139	151	136	134	122	130
ṣ	120	126	137	147	138	149	133	133	121	130
s	116	123	132	141	136	146	129	130	121	124
s	119	126	135	150	139	151	134	135	121	127
h	114	118	128	137	130	140	124	124	114	122
v	114	122	130	140	134	146	128	128	119	123
y	118	124	135	148	137	151	133	134	120	129
avg	116	122	132	143	135	146	129	130	119	125

C21. Effect of preceding syllable nucleus on present vowel

Sylla. nuclei	Present vowel									
	a	a:	i	i:	u	u:	e:	o:	ai	au
a	145	147	166	172	166	174	167	166	145	162
a:	142	144	159	168	166	167	156	159	142	159
i	150	152	166	175	169	178	167	163	145	153
i:	147	150	166	172	166	163	163	163	142	151
u	147	153	169	175	173	178	163	163	145	156
u:	145	150	162	172	169	178	156	156	142	151
e:	147	147	156	168	169	174	167	163	150	159
o:	145	147	166	172	173	178	160	166	150	165
ai	142	153	159	175	166	182	156	166	153	156
au	145	147	166	175	169	182	156	163	155	168
avg	146	149	164	172	169	175	161	163	147	158

C22. Effect of following syllable nucleus on present vowel

Sylla. nuclei	Present vowel									
	a	a:	i	i:	u	u:	e:	o:	ai	au
a	115	121	128	144	140	143	127	134	121	130
a:	120	119	129	144	135	146	133	132	120	127
i	120	119	132	137	139	151	128	132	112	126
i:	116	120	136	140	136	151	133	139	115	128
u	112	120	134	138	137	143	129	129	121	126
u:	119	126	136	140	135	141	133	130	124	123
e:	115	119	132	147	131	145	128	128	115	126
o:	112	120	137	145	135	145	133	129	116	128
ai	115	122	133	142	133	139	127	132	126	123
au	113	119	136	145	129	143	128	129	115	123
avg	116	121	133	142	135	145	130	131	119	126

REFERENCES

Allen, J. (1976). Synthesis of speech from unrestricted text. Proceedings of the IEEE, 64,433-442.

Allen, J. (1985). A perspective of man-machine communication by speech. Proceedings of the IEEE, 73, 1541-1551.

Allen, J., Hunnicutt, M. S. and Klatt, D. H. (1987). From Text-to-speech: The *MITalk* System, Cambridge University Press, Cambridge.

Akers, G. and Lennig, M. (1985). Intonation in text-to-speech synthesis: evaluation of algorithms. Journal of the Acoustical Society of America, 77, 2157-2165.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. Journal of the Acoustical Society of America, 52, 1687- 1697.

Atal, B. S. (1976). Automatic recognition of speakers from their voices. Proceedings of the IEEE, 64, 460-475.

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of speech wave. Journal of the Acoustical Society of America, 50, 637-655.

Atkinson, J. E. (1978). Correlation analysis of physiological factors controlling fundamental frequency. Journal of the Acoustical Society of America, 63, 211-222.

Brownston, L , Farrel, R., Kant, E. and Martin. M. (1986). Programming Expert *Systems* in *OOPS5*, Addison-Wesley, London.

Carpenter, G. A. and Grossberg, S. (1987). ART2: Self organization of stable category recognition codes for analog input patterns. Applied Optics, 26, 4919-4930.

Catford, J. C. (1988). A Practical Introduction to Phonetics, Clarendon Press, Oxford.

Chen, S. H. and Lin, M. T. (1987). On the use of pitch contour of Mandarin speech in text independent speaker identification. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Dallas, Texas, 3,1418-1421.

Childers, D. G. and Ke Wu. (1990). Quality of speech produced by analysis-synthesis. Speech Communication, 9, 97-117.

Childers, D. G., Ke Wu, Hicks, D. M. and Yegnanarayana, B. (1989). Voice conversion. Speech Communication, 8, 147-158.

Chomsky, N. and Halle, M. (1968). The Sound Pattern of English, Harper Row,

New York.

Cohen, A. and 't Hart, J. (1967). On the anatomy of intonation. Lingua, 19, 177-192.

Collier, R. (1975). Physiological correlates of intonation patterns. Journal of the Acoustical Society of America, 58,249- 255.

Cooper, W. E. and Paccia-Cooper, J. (1980). *Syntax* and Speech, Harward University Press, Cambridge.

Cooper, W. E. and Sorensen, J. M. (1977). Fundamental frequency contours at syntactic boundaries. Journal of the Acoustical Society of America, 62, 683-692.

Cooper, W. E. and Sorensen, J. M. (1981). Fundamental Frequency in Sentence Production, Springer-Verlag, New York.

Cooper, W. E., Soares, C., Ham, A. and Damon, K. (1983). The influence of inter- and intra-speaker tempo on fundamental frequency and palatelization. Journal of the Acoustical Society of America, 73,1723-1730.

Crystal, D. (1985). A Dictionary of Linguistics and Phonetics, Basil Blackwell, Cambridge.

Cutler, D.R. (1990). Exploiting prosodic probabilities in speech segmentation. In Cognitive *Models* of Speech Processing (Editor: Altmann, G.), MIT press, Cambridge.

Delgutte, B. (1978). A study of Perceptual investigation of F_0 contour with application to French. Journal of the Acoustical Society of America, 64, 1319-1332.

Dimshits, J. M. (1966). Hindi *vya:karaṇ ki: ru:p re:kha:*, Rajkamal Prakashan, New Delhi.

Dochery, G. and Shockey, L. (1988). Speech synthesis. In Aspects of Speech Technology (Editors: Jack, M. A. and Laver, J.), Edinburgh University Press, Edinburgh, 144-183.

Doddington, G. R. (1985). Identifying people from their voices. *Proceedings* of the IEEE, **73, 1651-1664.**

Ewan, W. G. (1979). Can intrinsic vowel F_0 be explained by source tract coupling? Journal of the Acoustical Society of America, **66, 358-362.**

Fant, G. (1960). Acoustic *Theory* of Speech Production, Mounton and Co., The Hague.

Fant, G. (1982). The voice source - acoustic modeling. Technical Report, STL-QPSR, **4/1982, 28-48.**

Fant, G., Kruckenberg, A. and Nord, L. (1991). Prosodic and segmental speaker variations. Speech *Communication*, **10,** 521-531.

Flanagan, J. L. (1972). Speech *Analysis Synthesis* and Perception, Springer-Verlag, New York.

Fujisaki, H. and Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. Journal of the Acoustic Society of Japan, (E)5, 233-242.

Fujisaki, H. and Kawai, H. (1988). Realization of linguistic information in the voice fundamental frequency contours of the spoken Japanese. In Proceedings of the International Conference on Acoustics, Speech *and* Signal Processing, New York, 2,663-666.

Garding, E. (1983). A generative model of intonation. In Prosody: *Models and* Measurements (Editors: Cutler, A. and Ladd, D. R.), Springer-Verlag, Berlin, 11-26.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. Cognitive Science, 11, 23-63.

Gussenhoven, C. and Rietveld, A. C. M. (1988). Fundamental frequency declination in Dutch: testing three hypotheses. Journal of Phonetics, 16, 355-369.

Haggard, M., Ambler, S. and Callow, M. (1970). Pitch as a voicing cue. Journal of the Acoustical Society of America, 47, 613-617.

Hair, G. D. and Rekeita, T. W. (1972). Mimic resistance of speaker verification using phoneme spectra. Journal of the *Acoustical* Society of America, 51, 131(A).

Harrington, J., Johnson, I. and Cooper, M. (1987). The application of phoneme sequence constraints to word boundary identification in automatic continuous speech recognition. In Proceedings of European Conference on Speech Technology, Edinburgh, England, 1,163-166.

Hayes-Roth, F., Waterman, D. A. and Lenet, D. B. (1983). Building *Expert* Systems, Addison-Wesley, London.

Hertz, S. R., Kadin, J. and Kerplus, K. J. (1985). A delta rule development system for speech synthesis from text. Proceedings of the IEEE, 73, 1589-1601.

Hess, W. (1983). Pitch Determination of Speech *Signals:* Algorithms *and Devices,* Springer-Verlag, Berlin.

Heinz, J. M. and Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *Journal* of the *Acoustical* Society of America, 33,589-596

Holmes, J. N. (1988). Speech Synthesis and Recognition, Van Nostrand Reinhold, England.

Kachru, Y. (1980). Aspects of *Hindi* Grammar, Manohar Publications, New Delhi.

Klatt, D. H. (1976). Structure of a phonological rule component for a synthesis-by-rule program. IEEE Transactions on Acoustics, Speech and Signal Processing, 24,391-398.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society *of America*, 67,971-995.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. Journal of the Acoustical Society of *America*, 82,737-793.

Kraayeveld, J., Rietveld, A. C. M. and Van Heauvan, V. J. (1991). Speaker characterization in Dutch using prosodic parameters. In Proceedings of the European Conference on Speech Communication, Genova, Italy, 1, 427-430.

Kohler, K. J. (1983). Prosodic boundary signals in German. Phonetica, 40, 89-134.

Kohler, K. J. (1990). Macro and micro F_0 in the synthesis of intonation. Papers in *Laboratory* Phonology I: Between Grammar and Physics of Speech (Editors: Kingston, J. and Beckman, M. E.), Cambridge University Press, Cambridge.

Kutik, E. J., Cooper, W. E. and Boyce, S. (1983). Declination of fundamental frequency in speakers production of parenthetical and main clauses. Journal of the Acoustical Society *of America*, 73,1731-1738.

Ladd, D. R. (1983). Peak features and overall slope. In Prosody: *Models* and Measurements (Editors: Cutler, A. and Ladd, D. R.), Springer-Verlag, Berlin, 39-52.

Ladd, D. R. (1988). Declination "reset" and hierarchical organization of utterances. Journal of the *Acoustical* Society *of America*, 84,530-544.

Ladd, D. R. and Silverman, K. E. A. (1984). Vowel intrinsic pitch in connected speech. Phonetica, 41, 31-40.

Ladd, D. R., Silverman, K. E. A., Tolkmit, F., Bergmann, G. and Scherer, K. R. (1985). Evidence for independent function of intonation contour type, voice quality and F_0 range in signaling speaker effect. Journal of the Acoustical Society of America, 78, 435-444.

Ladefoged, P. (1975). A Course in Phonetics, Harcourt Brace Jovanovich, New York.

Lea, W. A. (1980). Prosodic aids to speech recognition. In Trends in Speech Recognition (Editor: Lea, W. A.), Prentice-Hall, New Jersey, 166-205.

Lea, W. A., Medress, M. F. and Skinner, T. E. (1975). A prosodically guided speech understanding strategy. IEEE *Transactions* on Acoustics, Speech and Signal Processing, 23, 30-38.

Lee, S. H., Tseng, C. Y. and Young, M. O. (1989). Synthesis rules in a Chinese text-to-speech system. IEEE Transactions on Acoustic, Speech and Signal

Processing, 37, 1309-1320.
Lehiste, I. (1970). *Suprasegmentals*, M. I. T. Press, Cambridge.
Lehiste, I. and Peterson, G. E. (1959). Vowel amplitudes and phonemic stress in American English. Journal of the Acoustical Society of America, **31, 428-435.**
Lehiste, I. and Peterson, G. E. (1961). Some basic considerations in the analysis of intonation, Journal of the Acoustical Society of America, **33, 419-425.**
Levit, H. and Rabiner, L. R. (1971). Analysis of fundamental frequency contours in speech. Journal of the Acoustical Society of America, 49,569-582.
Lieberman, M. and **Pierrehumbert,** J. (1984). Intonation invariance under changes in pitch range and length; In Language Sound *Structure* (Editors: Aronoff, M and Oehrle, R. T.), M. I. T Press, Cambridge.
Lieberman, P. (1967). Intonation, Perception and Language, **MIT** Press, Cambridge.
Lieberman, P. and **Blumstein.** (1988). Speech Physiology, Speech Perception and Acoustic Phonetics, Cambridge University Press, Cambridge.
Lindau, M. (1986). Testing a model of intonation in a tone language. Journal of the Acoustical Society of America, **80, 757-** 764.
Lippmann, R. P. (1987). An introduction to computing with neural nets. IEEE Acoustics, Speech and *Signal* Processing Magazine, April **1987, 4-22.**
Lummins, R. C. 'and Rosenberg, A. E. (1972). Test of an automatic speaker verification method with intensively trained mimics. Journal of the Acoustical Society of America, **51,** 131(A).
Macchi, M., **Kahm,** C. and Streeter, L (1987). Expanding the template inventory for concatenative speech synthesis. In, Proceedings of the Speech *Technology'87,* **159-161.**
Maeda, S. (1974). A characterization of fundamental frequency contours of speech. *Quarterly Progress Report* of *Research* Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, **114, 1-14.**
Majewsky, W. and Blasdell, R. (1969). Influence of fundamental frequency cues on the perception of some synthetic intonation contours. Journal of the Acoustical Society of America, **45, 450-** 457.
Makhoul, J. (1975). Linear prediction: a tutorial review. Proceedings of the IEEE, 63,561-580.
Mariani, J. (1989). Recent advances in speech processing. In Proceedings of the International Conference on *Acoustic,* Speech and Signal Processing, Glasgow, Scotland, 1,429-440.
Markel, J. D. (1972). The SIFT algorithm for fundamental frequency estimation. IEEE *Transactions* on Audio and Electroacoustics, 20,367-377.

Mattingly, I. G. (1968). Experimental methods for speech synthesis by rule. IEEE Transactions on Audio and Electro Acoustics, 16, 198-202.

McClelland, J. L and Rumelhart, D. E. (1988). *Explorations* in Parallel Distributed Processing, The M I T Press, Cambridge.

Monaghan, A. I. C. and Ladd, D. R. (1990). Symbolic output as the basis for evaluating intonation in a text-to-speech system. Speech Communication, 9, 305-314.

Moore, R. R. (1965). A Study of Hindi Intonation, Ph. D Thesis, University of Michigan.

Nakagawa, S. and Sakai, T. (1979). A recognition system of connected spoken words based on word boundary detection. *Studia* Phonologica, University of Kyoto, Japan, 13.

Natakani, L. H. and Schaffer, J. A. (1978). Hearing "words" without words: prosodic cues for word perception, Journal of the Acoustical Society of America, **63, 234-245.**

Ohala, J. and Eukel, R. (1976). Explaining intrinsic pitch of vowels. Journal of the Acoustical Society of America, 60, **S44(A).**

Ohala, J. and Ewan, W. (1973). Speed of pitch range. Journal of the Acoustical Society of America, 53,345.

Ohala, M. (1983). Aspects of Hindi Phonology, Motilal Banarsidass, New Delhi.

Ohala, M. (1986). A search for the phonetic correlates of Hindi stress. In Structure, Convergence and Diglossia (Editor: Krishnamurthy, B), Motilal Banarsidas, New Delhi, 81-92.

Ohala, M. (1991). Phonological areal features of some Indo-Aryan languages. Language Sciences, 13,107-124.

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. Journal of the Acoustical Society of America, 75,224-230.

Ohman, S. E. G. (1966). Coarticulation in VCV utterances: spectrographic measurements. Journal of the Acoustical Society of America, **39, 15**1-168.

Olive, P. J. (1975). Fundamental frequency rules for the synthesis of simple declarative English sentences. Journal of the Acoustical Society of America, **57, 476-482.**

Oppenheim, A. V. and Schaffer, R. W. (1975). Digital Signal Processing, Prentice-Hall, New Jersey.

O'Shaughnessy, D. (1976). Modeling Fundamental Frequency and *Its* Relationship to *Syntax,* Semantics and Phonetics, Ph D Thesis, Massachusetts Institute of Technology, Cambridge.

O'Shaughnessy, D. (1984). Design of a real-time French text-to-speech system.

Speech Communication, 3, 233-243.
O'Shaughnessy, D. (1986). Speaker recognition. IEEE Acoustic, Speech and Signal Processing Magazine, October 1986, 4-27.
O'Shaughnessy, D. (1987). Speech communication: human and machine. Addison Wesley, Massachusetts.
O'Shaughnessy, D. and Allen, J. (1983). Linguistic modality effect on fundamental frequency in speech. Journal of the Acoustical Society of America, 74, 1155-1172.
Pandey, P. K. (1989). Word accentuation in Hindi. Lingua, 77, 37- 73.
Papamichalis, P. E. (1987). Practical Approaches to Speech *Coding*. Prentice-Hall, New Jersey.
Peterson, G. E. and Barney, H. L (1952). Control methods used in a study of vowels. Journal of the Acoustical Society of America, 24, 175-184.
Peterson, N. (1978). Intrinsic fundamental frequency of Danish vowels. Journal of Phonetics, 6, 177-189.
Picket, J. M. (1980). The Sound of Speech Communication. University Park, Baltimore.
Pierrehumbert, J. (1979). Perception of fundamental frequency declination. Journal of the Acoustical Society of America, 65,363- 369.
Pierrehumbert, J. (1981). Synthesizing intonation. Journal of the Acoustical Society of America, 70,985-995.
Pisoni, D. B., Nusbaum, H. C. and Green, B. G. (1985). Perception of synthetic speech generated by rule. Proceedings of the IEEE, 73, 1665-1676.
Pollack, I., Pickett, J. M. and Samby, W. H. (1954). On talker identification by their voice. *Journal* of the *Acoustical* Society of America, 26, 403-406.
Quene, H. and Kager, R. (1992). The derivation of prosody for text-to-speech from prosodic sentence structure. Computer Speech and *Language*, 6, 77-98.
Rabiner, L. R., Cheng, M. J., Rosenberg, A. E. and McGonegal, C. A. (1976). A comparative performance study of several pitch detection algorithms. IEEE Transactions on Acoustic, Speech *and* Signal Processing, 24, 399-418
Rabiner, L. R., Levitt, H. and Rosenberg, A. E. (1969). Investigations of stress patterns for speech synthesis by rule. Journal of the Acoustical Society of *America*, 45, 92-101.
Rabiner, L. R. and Schaffer, R. W. (1978). Digital Processing of Speech Signals, Prentice-Hall, New Jersey.
Rajesh Kumar, S. R. (1990). Significance of Durational *Knowledge* in a Text-to-speech System for Hindi, M. S. Thesis, Indian Institute of Technology, Madras.

Ramachandran, V. R. (1992). Coarticulation Rules for a Text-to-speech System for Hindi, M. S. Thesis, Indian Institute of Technology, Madras.

Ramana Rao, G. V. and Yegnanarayana, B. (1991). Word boundary hypothesization in Hindi speech. Computer Speech and language, 5, 379-392

Reich, A. and Duke, J. (1979). Effect of selected vocal disguises upon speaker identification by listening. Journal of the Acoustical Society of America, 66, 1023-1028.

Reich, A., Moll, K. and Curtis, J. (1976). Effects of selected vocal disguises upon spectrographic speaker identification. Journal of the Acoustical Society of America, 60, 919-925.

Rich, E. (1983). *Artificial* Intelligence. McGraw Hill, Singapore.

Rietveld, A. C. M. and Gussenhoven, C. (1987). Perceived speech rate and intonation. Journal of Phonetics, 15, 273-285.

Rogers, D. F. and Adams, J. A. (1989). *Mathematical* Elements for Computer Graphics, McGraw Hill, New York.

Rosenberg, A. E. (1968). Effect of pitch averaging on the quality of natural vowels. *Journal* of the Acoustical Society of *America*, 44, 1592-1595.

Rosenberg, A. E. (1976). Automatic speaker verification: a review. Proceedings of the IEEE, 64,475487.

Rumyanceva, I. M. (1988). Hindi word prosody (Experimental research). In Linguistics: A Soviet approach (Editors: Andronov and Mallik, B. P), Indian Journal of Linguists, Calcutta, 395- 431.

Sarma, V. V. S. and Yegnanarayana, B. (1975). A critical survey of automatic speaker recognition systems. Journal of Computer Society of India, 6, 1-11.

Sato, H. (1984). Japanese text-to-speech conversion system. Review of Electrical Communication Laboratory, Nippon Telegraph and Telephone Corporation, 32, no:2.

Shaddle, C. H. (1985). Intrinsic fundamental frequency of vowels in sentence context. *Journal* of the Acoustical Society of America, 78, 1562-1567.

Sharma, A. (1969). Hindi word accent. Indian Linguistics, 30, 115-8.

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. Journal of the Acoustical Society of America, 64, 1582- 1591.

Terken, J. and Lameer, G. (1988). Effect of segment quality and intonation on quality judgments for texts and utterances. Journal of Phonetics, *16*, 453-457.

't Hart, J., Collier, R. and Cohen, A. (1990). A Perceptual Study of Intonation. Cambridge University Press, Cambridge.

Thorsen, N. G. '(1980). A study of perception of sentence intonation - evidence

from Danish. Journal of the Acoustical Society of America, 67, 1014-1030.
Thorsen, N. G. (1983). Two issues in the prosody of standard Danish. In Prosody: Models and Measurements (Editors: Cutler, A. and Ladd, D. R.), Springer-Verlag, Berlin, 27-38.
Thorsen, N. G. (1985). Intonation and text in standard Danish. Journal of the Acoustical Society *of America*, 77,1205-1216.
Thorsen, N. G. (1986). Sentence intonation in textual contexts - Supplementary data. Journal of the Acoustical Society of America, 80, 1041-1047.
Titz, I. R. (1989). Physiologic and acoustic differences between male and female voices. Journal of the Acoustical Society of America, 85,1699-1707.
Umeda, N. (1976). Linguistic rules for text-to-speech synthesis. Proceedings of the IEEE, 64,443-451.
Umeda, N. (1981). Influence of segmental factors on fundamental frequency of fluent speech. Journal of the Acoustical Society of America, 70,350-355.
Umeda, N. (1982). F_0 declination is situation dependent. Journal of Phonetics, 10,279-290.
Van Bezooijen, R. qnd Pols, L. C. W. (1990). Evaluating text-to- speech systems: some methodological aspects. Speech Communication, 9,263-270.
Vanden Berg, J. (1968). Mechanism of larynx and laryngeal vibrations. In Manual of Phonetics (Editor: Malmberg, B.), North Holland, Amsterdam, 278-308.
Vaissiere, J. (1974). On French prosody. *Quarterly* Progress *Report from* Research Lab of Electronics, Massachusetts Institute of Technology, 114,212-223.
Vaissiere, J. (1983). Language independent prosodic features. In Prosody: *Models and* Measurements (Editors: Cutler, A. and Ladd, D. R.), Springer-Verlag, Berlin.
Waibel, A. (1986). Suprasegmental in very large vocabulary word recognition. In Pattern Recognition by Humans and Machines: Speech Perception, Academic Press, London, 1,159-186.
Waibel, A. (1987). Prosodic knowledge source for word boundary hypothesization in a connected speech recognition system. In Proceedings of the *International* Conference on Acoustics, Speech and Signal Processing, Dallas, Texas, 2,856-889.
Waibel, A. (1988). Prosody and Speech Recognition, Morgon Kauffman, London.
Wang, M. Q. and Hirschberg, J. (1992). Automatic classification of intonation phrase boundaries. Computer Speech and Language, 6, 175-196.
Wasserman, P. D. (1989). Neural Computing Theory and Practice, Van Nostrand Reinhold, New York.

Wightman, C. W. and Ostendorf, M. (1991). Automatic recognition of prosodic phrases. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing,* Toronto, Canada, 1,321- 324.

Willems, E. E. and Stevens, K. N. (1972). Emotions and speech: some acoustic correlates. *Journal of the Acoustical Society of America,* 52,1238-1249.

Willems, W., Collier, R. and 't Hart, J. (1988). A synthesis scheme for British English intonation. *Journal of the Acoustical Society of America, 84,* 1250-1261.

Witten, I. H. (1982). *Principles of Computer Speech,* Academic Press, London.

Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition, *Journal of the Acoustical Society of America,* 51,2044-2055.

Yegnanarayana, B. and Murthy, Hema. A. (1988). Formant extractions using group delay functions. In *Proceedings of the International Workshop on Speech Processing,* Bombay, 31-41.

Yegnanarayana, B., Murthy, Hema. A. and Ramachandran, V. R. (1991). Speech processing using modified group delay functions. In *Proceedings of the International Conference on Acoustic, Speech and Signal Processing,* Toronto, Canada, 2,945-948.

Yegnanarayana, B., Murthy, Hema. A., Sundar, R., Alwar, N., Ramachandran, V. R., Madhukumar, A. S. and Rajendran, S. (1990). Development of a text-to-speech system for Indian languages. In *Frontiers of knowledge based computing systems* (Editors: Rege and Bhatkar), Narosa Publishing house, Bombay, 467-476.

Yegnanarayana, B. and Ramachandran, V. R. (1992). Group delay processing of speech signals. In *Proceedings of the ESCA Workshop on "Comparative speech signal representations",* Sheffield, England, 411-418.

Young, S. J. and Fallside, F. (1979). Speech synthesis from concepts: a method for speech output from information systems. *Journal of the Acoustical Society of America,* 66,685-695.

Zawadsky, P. A. and Gilbert, H. R. (1989). Vowel fundamental frequency and articulator position. *Journal of Phonetics,* 17,159-166.

Zue, V. M. and Schartz, R. M. (1980). Acoustic procesing and phonetic analysis. In *Trends in Speech Recognition* (Editor: Lea, W. A), Prentice-Hall, New Jersey, 101-124.

LIST OF FIGURES

Fig.3.6. Speech waveform and F_0 contour for yes-no type interrogative sentences

(a) */kya: ga:ḍi: samay par hail* (Whether train is on time)

(b) */kya: vah duka:n band thi:/* (Whether the shop was closed)

The falls and rises of F_0 contour rise towards the end of the utterance.

Fig.3.7. Speech waveform and F_0 contour for question-word type interrogative sentences

(a) */ajay ke: pa:s kaun si: kita:b hail* (Which book is with Ajay)

(b) */kalkatte: mẽ: tumne: kya: dekha:/* (What have you seen in Calcutta)

The top line and the base line decline gradually up to the question word and then rise towards the end.

Fig.3.8. Pitch accent patterns for monosyllabic words

(a) Monosyllabic content word */ham/* (we)

(b) Monosyllabic function word */jo:/* (that)

(c) Monosyllabic function word */ko:/* accented with the previous content word */manuṣy/* (man)

(d) Monosyllabic function word */ki:/* conjoined with the previous content word */j:van/* (life)

In (a) and (b), the valley and the peak of F_0 contour occur in the same syllable. In (c), the peak of F_0 contour of the previous content word is shifted to the function word. In (d), F_0 contour of the function word decreases monotonously.

Fig.3.9. Pitch accent patterns for disyllabic words

(a) */pra:rtna:/* (prayer)

(b) */ke:val/* (absolute)

(c) */ja:grit/* (awake)

In all the cases valley occurs on the initial syllable and the peak occurs on the final syllable. In (b), the peak of F_0 contour is shifted to the coda since the coda is lateral.

Fig.3.10. Pitch accent patterns for trisyllabic words

(a) */tu:pha:ni:/* (storm)

(b) */pariṇa:m/* (consequence)

(c) */gali: ke:/* (street)

In all the cases valley occurs on the initial syllable and the peak occurs on the final syllable. In (b), the peak of F_0 contour is shifted to the coda since the coda is nasal.

Fig.3.11. Pitch accent patterns for tetrasyllabic words

(a) */sahiṣṇuta:/* (tolerance)

(b) */antarya:mi:/* (omniscient)

The word (a) is monomorphemic and correspondingly the valley occurs on the initial syllable and the peak occurs on the final syllable. The word (b) is bimorphemic and hence the F_0 contour is characterized by two valleys and two peaks in alternate syllables.

Fig.3.12. Pitch accent patterns for pentasyllabic words

(a) */samajhne:va:la:/* (one who understands)

(b) */sarvaśaktima:n/* (almighty)

Both the words are polymorphemic and correspondingly two valleys and two peaks occured in the Fo contour.

Fig.3.13. Pitch accent patterns for the prosodic word **(*/mi:na: ko:/*)** when it occurs in the following three positions in a sentence.

(a) */mi:na: ko: sita:r usta:d ne: sikha:ya:/*

(b) */sita:r mi:na: ko: usta:d ne: sikha:ya:/*

(c) */usta:d ne: mi:na: ko: sita:r sikha:ya:/*

(Master taught sitar to Mina)

The valleys and the peaks of Fo contour of a prosodic word does not change even if the word occurs at different positions in a sentence. But the range of F_0 contour changes according to the position.

Fig.4.1. Speech waveform and F_0 contour for complex declarative sentences

(a) */jo: kita:b tumne: mujhko: di: thi: vah śya:m ke: pa:s hai/* (The book which you gave me is with **Syam)**

(b) */ra:m ne: idhar ca:y piya: udhar kamre: mẽ? a:ra:m kiya:/* (Ram took tea from here and took rest in that room)

Each sentence has two syntactic clauses separated by a pause (P). Resetting of F_0 contour occurs at the beginning of a new syntactic clause.

Fig.4.2. Changes of F_0 contour due to the presence of negation

(a) Fo contour and the utterance for the sentence */mãĩ ka:ṇpu:r ja: raha: hũ:/* (I am going to Kanpur)

(b) F_0 contour and the utterance for the sentence */mãĩ ka:ṇpu:r nahĩ: ja: raha: hũ:/* (I am not going to Kanpur)

The valleys and the peaks Fo contour in the negation word */nahĩ:/* is emphasized in (b).

Fig.4.3. Changes of F_0 contour due to the presence of numerals

(a) */ra:m ki: mo:tar sa:i:kil te:i:s haja:r che: sau aṭha:rah ki: hail* (Ram's

motor cycle costs twenty three thousand six hundred and eighteen.

(b) *lye: kita:b painta:lis rupaye: Id: hail* (This book costs forty five rupees)

F_0 contour resets at the beginning of the numeral and the magnitude of resetting is proportional to the number of words in the numeral.

Fig.5.1. Inherent F_0 of vowels. Inherent F_0 of a vowel is proportional to the quality and the quantity of the vowel. High vowels exhibit higher Fo than low vowel. Inherent F_0 of longer vowels are more compared with their shorter counter parts.

Fig.5.2. Effect of the preceding consonant on inherent F_0 of the vowels */a:/* and */u:/*

(a) When consonants are classified based on the voicing nature and manner of articulation

(b) When consonants are classified based on the place of articulation

Straight line indicates the average F_0 of the corresponding vowel

Fig.5.3. Effect of the following consonant on inherent F_0 of the vowels */a:/* and */u:/*

(a) When consonants are classified based on the voicing nature and manner of articulation

(b) When consonants are classified based on the place of articulation

Straight line indicates the average F_0 of the corresponding vowel

Fig.5.4. Effect of the adjacent syllables on inherent F_0 of the vowels */a:/* and */u:/*

(a) Changes in inherent F_0 due to the change in the nucleus of the preceding syllable

(b) Changes in inherent F_0 due to the change in the nucleus of the following syllable

Fig.6.1. Block diagram of a text-to-speech system for Hindi

Fig.6.2. Block diagram for a basic speech production model. The vocal tract system is represented by LPCs and formants. Pitch period and gain are the source parameters. The voiced source of the speech is represented by an excitation signal and the unvoiced source is represented by random noise.

Fig.6.3. Incorporation of intonation knowledge in a text-to- speech for Hindi

Fig.6.4. Block diagram of an intonation parser to determine the sentence types and to assign appropriate F_0 contour.

Fig.6.5. Block diagram for the word analyzer. Type 1 function word corresponds to the function words which have independent existence. Type 2 corresponds to unaccented function words and Type 3 corresponds to

accented function words.

Fig.6.6. Block diagram of the character analyzer used in our text-to-speech system

Fig.6.7. Output of the text processing modules in our text-to-speech system. CW indicates content word and FW2 indicates unaccented function word. UWA, UVA and W A indicate unvoiced unaspirated stop, unvoiced aspirated stop and voiced aspirated stop, respectively.

Fig.6.8. Speech waveform and F_0 contour for a simple declarative sentence */śaṅkar ja:ta: hail* (Shankar goes)

(a) Natural speech signal

(b) Synthesized speech signal without applying intonation knowledge

(c) Synthesized signal after applying intonation knowledge

Fig.6.9. Speech waveform and F_0 contour for an yes/no type interrogative sentence */kya: śaṅkar pa:s hogaya:?/* (Has Shankar passed?)

(a) Natural utterance

(b) Synthesized speech

Fig.6.10. Speech waveform and F_0 contour for a question-word type interrogative sentence */tum kahã: se: ate: ho:?/* (Where did you come from?)

(a) Natural utterance

(b) Synthesized speech

Fig.6.11. Speech waveform and F_0 contour for a complex declarative sentence */a:tma: amar hai śari:r na:śva:n hail* (Soul is immortal, body is mortal)

(a) Natural utterance

(b) Synthesized speech

Fig.7.1. Block diagram of a speech-to-text system for Hindi

Fig.7.2. Examples of pitch accent features in Hindi

(a) Disyllabic word */ke:val/* (absolute) (LH)

(b) Monomorphemic tetrasyllabic word */sahiṣṇuta:/* (tolerance) $(LH_1H_2H_3)$

(c) Two disyllabic word with an accented function word in between */pra:rtna:#to:#a:tma:/* (LHL_1L_2H). # indicates a word boundary.

Fig.7.3. Algorithm for hypothesizing word boundaries from continuous speech in Hindi

Fig.7.4. Hypothesization of word boundaries for the natural utterance corresponding to the sentence */śaṅkar#ja:ta:#hai/* (Shankar goes). Thick line indicates the pitch contour for the utterance and thin line corresponds to the energy contour. Vertical bars in energy contour indicate the location

of syllable nuclei. # indicates a word boundary. L and H indicate pitch accent feature of each syllable.

Fig.7.5. Hypothesization of word boundaries for adverse speech input conditions. The clean speech corresponding to the sentence */śaṅkar#ja:ta:#hai/* (Shankar goes) is mixed with random noise **(SNR = 3dB).** Thick line indicates the pitch contour for the utterance and thin line corresponds to the energy contour. No **voiced/unvoiced** decision is taken at the pitch estimation. Vertical bars in energy contour indicate the location of syllable nuclei. # indicates a word boundary. L and H indicate pitch accent feature of each syllable.

Fig.7.6. Algorithm for hypothesizing function words in Hindi from continuous speech using word boundary **hypothesization** algorithm

Fig.7.7. Hypothesization of word boundaries for continuous speech corresponding to the following Hindi sentences.

(a) */pra:rtna: ko: a:tma: ko: sa:f karne: ka: ja:ḍu: hail*

(b) */jo: kisi: se: i:rṣya nahĩ: rahta: jo: daya: ka: bhanḍa:r hai/*

(c) */a:tma śudhi: ka: arth ji:van ki: pahalõ: se: śudhi: ho:na: ca:hiye:/*

(d) */har e:k manuṣya ke: andar se: i:śvar bo:lta: hai hi:/*

(e) */sab dharm e:k hi: stha:n par pahũ:c ne: ke: alag alag raste: hail*

The algorithm detected word boundaries and located function words. Here, P indicates pause corresponding to syntactic boundaries, # indicates the word boundaries hypothesized by the algorithm, • indicates actual word boundaries and fw corresponds to the hypothesized function words.

Fig.8.1. Block diagram of the proposed speaker recognition system. It consists of two parts, one corresponds to the feature extraction and the other for classification. The features are pitch frequency values at valleys and peaks of pitch contour and durations of the words in the utterance of carrier sentence. Words are hypothesized using word boundary hypothesization algorithm with the help of pitch and energy contour. The classifier is a neural network based system in which neural network model is selected based on the domain of application.

Fig.8.2. Hypothesization of word boundaries using word boundary hypothesization algorithm for the test utterance of the carrier sentence */bha:rat#hama:ra:#de:ś#hai/* (India is our country). Thick line indicates the pitch contour for the utterance and the thin line corresponds to the energy contour. Vertical bars in energy contour indicate the locations of

the syllable nuclei. Circles marked in the pitch contour correspond to pitch frequency value at the midpoints of the syllable nuclei. # indicates a word boundary hypothesized by the algorithm.

Fig.8.3. Variation of pitch contour with different noise levels. Clean speech signal is mixed with white noise at signal to noise ratio (SNR) levels of 20dB, 10dB, 5dB and 3dB. Pitch contours are extracted using the properties of group delay functions. Plots of pitch contours for two speakers are given.

Fig.8.4. Hypothesization of word boundaries for noisy speech input conditions. The clean speech corresponding to the carrier sentence */bha:rat# hama:ra:#de:ś#hai/* (India is our country) is mixed with white noise (SNR = 3dB). Thick line indicates the pitch contour for the utterance and the thin line corresponds to the energy contour. Vertical bars in energy contour indicate the locations of the syllable nuclei. Circles marked in the pitch contour correspond to pitch frequency value at the midpoints of the syllable nuclei. # indicates a word boundary hypothesized by the algorithm.

Fig.8.5. Architecture of a neural network based on back propagation algorithm

(a) Model of a basic neuron element.

(b) Architecture of the proposed speaker recognition system based on back propagation algorithm. The network consists of two layers, one hidden layer and an output layer. Training of this network involves two phases. Thick line indicates the forward phase of training, in which output of each neuron is computed. Dotted line corresponds to the backward propagation of errors where errors are determined as the differences between the desired and actual outputs.

Fig.8.6. Number of hidden nodes and the number of training cycles for a network.

(a) Variation of mean squared error with respect the number of hidden nodes. Number of epochs for the training is taken as constant (2000epochs).

(b) Variation of mean square error with respect to the number of epochs. Number of hidden nodes is taken as constant (20).

Fig.8.7. Architecture of a neural network based on adaptive resonance theory. F1 and F2 are feature representation and category representation field, respectively, analogous to the short term memory (STM) of brain. LTM, long term memqry encodes new input patterns as the traces of bottom-up

and top-down adaptive filter representations. These together called the attentional subsystem of the network. Gainl, gain2 and vigilance are the parameters of orienting subsystem which helps for the stabilization of the network (Carpenter & Grossberg, 1987).

Fig.8.8. Improvement in the performance of speaker recognition by considering pitch and durational features together.
(a) In the neural network based on back propagation algorithm
(b) In the neural network based on adaptive resonance theory

Fig.8.9. Confusion matrix obtained for the speaker recognition in clean speech.
(a) Using back propagation algorithm
(b) Using adaptive resonance theory
Rows in the matrix indicates the actual speaker and the columns indicate the identified speaker. Column X corresponds to the unrecognized speech utterances. The accuracy of speaker recognition for clean speech using BP algorithm is 96.25% and using ART algorithm is 86.88%

Fig.8.10. Performance degradation in the speaker recognition due to the addition of noise in speech. Error rates for a BP based system are 3.75%, 4.38%, 5.63%, 10% and 15% for clean speech, and noisy speech signals of SNR 20dB, 10dB, 5dB and 3dB, respectively. Similarly, the corresponding error rates for ART based system is 13.12%, 13.12%, 12.50%, 16.88% and 25%.

LIST OF TABLES

LIST OF PUBLICATIONS

1. Rajesh Kumar, S. R., Ramachandran, V. R., Madhukumar, A. S., Murthy, Hema A. and Yegnanarayana, B. (1990). A text-to-speech system for Indian languages. Presented at the Seminar on Common Phonetic *Matrix* for Indian *Languages*, Central Institute for Indian Languages, March 1990, Mysore, India.
2. Yegnanarayana, B., Murthy, Hema A., Sundar, R., Alwar, N., Ramachandran, V. R., Madhukumar, A. S. and Rajendran, S. (1990). Development of a text-to-speech system for Indian Languages. In Frontiers in knowledge-based computing (Editors: Bhatkar, V. P. and Rege, K. M.), Narosa Publishers, New Delhi, 467 - 476.
3. Madhukumar, A. S., Rajendran, S. and Yegnanarayana, B. (1991). Significance of prosodic knowledge in a text-to-speech system for Hindi. In Proceedings of the *XII* International Congress on Phonetic Sciences, *Aix-en-provence*, France, 3,494-497.
4. Madhukumar, A. S., Rajendran, S., Chandra Sekhar, C. and Yegnanarayana, B. (1991). Synthesizing Intonation for Speech in Hindi. In Proceedings of the 11 European Conference on Speech Communication and *Technology*, Genova, Italy, 3, 1153-1156.
5. Madhukumar, A. S., Rajendran, S. and Yegnanarayana, B. (1991). Synthesizing intonation for Indian Languages. In Proceedings of the *VI* Annual International Conference of IEEE Region 10, New Delhi, India, 2, 224-228.
6. Yegnanarayana, B., Rajendran, S., Rajesh Kumar, S. R., Ramachandran, V. R. and Madhukumar, A. S. (1992). Knowledge sources for a text-to-speech system in Hindi. In Computer Processing of Asian Languages (Editor: Sinha. R. M. K), Tata McGraw-Hill Publishing Company, New Delhi, 233-242.

7. Yegnanarayana, B,, Madhukumar, A. S. and Ramachandran, V. R. (1992). Robust features for applications in speech and speaker recognition. In Proceedings of the European Speech Communication Association Workshop on Speech processing in adverse *conditions*, Cannes Mandelieu, France, November 1992.

8. Madhukumar, A. S. and Yegnanarayana, B. (1993). A speaker recognition system based on intonation knowledge. In Proceedings of the *Discussion* meeting on recent advances in *Signal* Processing and Communications, Bangalore, India, January 1993.

9. Madhukumar, A. S., Rajendran, S. and Yegnanarayana, B. Intonation component of a text-to-speech system for Hindi. Accepted for publication, Computer Speech and *Language*, Academic Press, London.

10. Rajendran, S., Madhukumar, A. S. and Yegnanarayana, B. Word boundary hypothesization for continuous speech in Hindi based on pitch accent patterns. Communicated to Computer Speech *and Language*, Academic Press, London.

11. Madhukumar, A. S. and Yegnanarayana, B. Neural networks for speaker recognition based on intonation knowledge. Communicated to IEEE Transactions on Neural Networks.